500.43154X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): Y. OGAWA, et al.

Serial No.: 10/671,718

Filed: September 29, 2003

Title: METHOD AND DEVICE FOR RELEVANT DOCUMENT SEARCH

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

November 6, 2003

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on:

Japanese Patent Application No. 2003-089633
Filed: March 28, 2003

A certified copy of said Japanese Patent Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge
Registration No.: 29,621

CIB/rr
Attachment

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 Date of Application:	２００３年　３月２８日
出願番号 Application Number:	特願２００３－０８９６３３
[ST. 10/C]:	[ＪＰ２００３－０８９６３３]
出願人 Applicant(s):	株式会社日立製作所

２００３年１０月　２日

特許庁長官
Commissioner,
Japan Patent Office　今井康夫

出証番号　出証特２００３－３０８１１９９

【書類名】　特許願
【整理番号】　K02011981A
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
　【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内
　【氏名】　小川　祐一
【発明者】
　【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内
　【氏名】　松林　忠孝
【発明者】
　【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内
　【氏名】　山本　伸也
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100075096
　【弁理士】
　【氏名又は名称】　作田　康夫
【手数料の表示】
　【予納台帳番号】　013088
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】

類似文書検索方法および類似文書検索装置

【特許請求の範囲】

【請求項１】

検索対象の文書（以下、対象文書という）の中から文書を検索する類似文書検索方法であって、

検索条件として入力された全文検索条件を取得し、

前記対象文書を複数の部分に分割し、

該分割した前記対象文書の各部分に対して前記全文検索条件に対する類似度を算出し、

該算出した類似度と予め定められた閾値とを比較して、前記分割された各部分が前記全文検索条件に適合している部分であるか否かを判定し、

該判定結果をもとに、前記分割された部分を含む対象文書の前記全文検索条件に対する詳細度を算出することを特徴とする類似文書検索方法。

【請求項２】

前記類似文書検索方法はさらに、

前記対象文書の前記全文検索条件との類似度を算出し、

該算出した前記全文検索条件に対する類似度と、前記算出した前記対象文書の前記全文検索条件に対する詳細度とを表示することを特徴とする請求項１記載の類似文書検索方法。

【請求項３】

検索対象の文書（以下、対象文書という）の中から関連する文書を検索する類似文書検索装置において、

検索条件である種文書を取得する種文書取得手段と、

前記対象文書を複数の部分に分割する分割手段と、

前記取得された種文書をもとに、該分割された前記対象文書の各部分について前記種文書に対する類似度を算出する類似度算出手段と、

該算出された部分ごとの類似度が、所定の値を超えているか否かを判断して、

前記対象文書の前記種文書に対する詳細度を算出する詳細度算出手段とを備えることを特徴とする類似文書検索装置。

【請求項４】

前記類似文書検索装置はさらに、

前記対象文書に対する全文検索における全文検索条件を解析する全文検索条件解析手段と、

該解析された全文検索条件をもとに、前記分割された各部分に対して全文検索を行い、前記各部分の前記全文検索条件に対する適合度を算出する全文検索条件適合度算出手段とを備え、

前記詳細度算出手段は、該全文検索条件適合度と、前記類似度算出手段が算出した前記分割された各部分について前記種文書に対する類似度とを用いて、前記対象文書の前記種文書に対する詳細度を算出することを特徴とする請求項３記載の類似文書検索装置。

【請求項５】

前記類似文書検索装置はさらに、

前記種文書に対する詳細度または前記種文書に対する類似度をキーとして、検索対象である複数の対象文書に対して順位をつけて表示する表示手段を備えることを特徴とする請求項３記載の類似文書検索装置。

【請求項６】

検索対象の文書（以下、対象文書という）の中から関連する文書を検索する類似文書検索プログラムを格納した記憶媒体であって、

前記対象文書を検索する検索条件である種文書を取得するステップと、

前記対象文書を複数の部分に分割するステップと、

前記取得された種文書をもとに、前記分割された各部分の前記種文書に対する類似度を算出するステップと、

該算出された類似度を、所定の閾値と比較するステップと、

該比較の結果を用いて、前記分割された各部分が前記種文書に適合しているか否かを判定して、前記種文書に適合している部分の数を集計するステップと、

該集計した数をもとに、前記対象文書の前記種文書に対する詳細度を算出する

ステップとを備えることを特徴とする類似文書検索プログラムを格納した記憶媒体。

【請求項７】

予め記憶された検索対象の文書（以下、対象文書という）と、検索条件である文書（以下、種文書という）との関連性を判定する文書間関連性判定方法であって、

前記対象文書を複数の部分に分割し、

該分割した前記対象文書の各部分に対して前記種文書に対する類似度を算出し、

該算出した類似度と予め定められた閾値とを比較して、前記分割された各部分が前記種文書に適合している部分であるか否かを判定し、

該判定結果をもとに、前記種文書に適合している部分の数を集計し、

該集計した部分の数をもとに、前記分割された部分を含む対象文書の前記種文書に対する詳細度を算出することを特徴とする文書間関連性判定方法。

【請求項８】

前記文書間関連性判定方法はさらに、

前記対象文書の前記種文書に対する類似度を算出し、

該算出した前記種文書に対する対象文書の類似度と、前記算出した前記種文書に対する対象文書の詳細度の少なくとも一方を出力することを特徴とする請求項７記載の文書間関連性判定方法。

【請求項９】

前記文書間関連性判定方法はさらに、

前記対象文書に対して全文検索を行う場合の全文検索条件を取得し、

該取得した全文検索条件を用いて、前記分割された対象文書の各部分に対して全文検索を行い、前記分割された対象文書の部分ごとに前記全文検索条件に対する適合度を算出し、

該算出した前記全文検索条件に対する適合度と、前記算出した対象文書の部分ごとの前記種文書に対する類似度とを用いて、対象文書の中で前記検索に適合する部分を検出することを特徴とする請求項７記載の文書間関連性判定方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ユーザが指定した文書間の類似性の判定指標を算出する文書間関連度算出方法とこれを用いた類似文書検索方法に関する。

【０００２】

【従来の技術】

近年、パーソナルコンピュータやインターネットの普及に伴い、電子化された文書が大量に存在するようになった。その大量の文書の中からユーザーが目的とする文書を効率よく検索する文書検索技術が盛んに開発されており、中でも検索条件として入力された文書（以下、種文書と呼ぶ）と類似した文書を検索する類似文書検索が注目されている。

【０００３】

この類似文書検索に関して、特開平９－１６０９２８号公報には、種文書を構成する文と、種文書に対する類似度を算出する文書（以下、対象文書と呼ぶ）を構成する文の全組み合わせに対して文間の類似度を算出し、それらの類似度を加算することで文書全体の類似度を算出する技術が開示されている。例えば、種文書がＡ、Ｂの２文で構成され、対象文書がＣ、Ｄ、Ｅの３文で構成されている場合、種文書に関する対象文書の類似度は、（ＡとＣの類似度）、（ＡとＤの類似度）、（ＡとＥの類似度）、（ＢとＣの類似度）、（ＢとＤの類似度）、（ＢとＥの類似度）の和として算出される。これにより、種文書に関する内容が対象文書の全体で類似している場合に高い類似度の値が算出される。

【０００４】

【特許文献１】　特開平９－１６０９２８号公報

【発明が解決しようとする課題】

しかし、上記従来技術では、ある文間の類似度が極端に高い場合、他の文間の類似度が低くても文書全体の類似度としては高くなってしまう場合がある。すなわち、ある対象文書に対して高い類似度が算出された場合、対象文書の全体が類似している場合と対象文書の一部が類似している場合が考えられる。検索者はこ

れらの違いを区別できないため、ユーザは目的に応じた種文書に関する効率的な検索が行なえない。例えば、種文書に記載された内容に関して幅広く情報を得るために文書全体で類似している対象文書を参照したい場合、上記従来技術を用いて算出された類似度では判断できない。

【０００５】

本発明の目的は、文書の類似性を判断するための指標を提示する類似文書検索方法を提供することにある。

【０００６】

【課題を解決するための手段】

上記目的を達成するために本発明は、予め記憶された検索対象文書の中から文書を検索する検索条件として入力された種文書に含まれる文字列を抽出し、対象文書を複数の部分に分割して、分割した対象文書の各部分に含まれる文字列を抽出し、これら文字列を比較して、前記分割された部分ごとに前記種文書に対する類似度を算出するとともに、その類似度と予め定められた閾値とを比較して、分割された各部分が種文書に適合している部分であるか否かの判定結果をもとに、対象文書の前記種文書に対する詳細度を算出する構成を採用した。

【０００７】

【発明の実施の形態】

以下に、本発明の第一の実施例について説明する。

【０００８】

図１は、本実施例で示す文書検索システムの全体構成図を示す。本システムは、ディスプレイ１００、キーボード１０１、中央演算処理装置（CPU）１０２、磁気ディスク装置１０３、フレキシブルディスクドライブ（ＦＤＤ）１０４、主メモリ１０５、これらを結ぶバス１０６および他の機器と本システムを接続するネットワーク１０７から構成される。

【０００９】

磁気ディスク装置１０３は二次記憶装置の一つであり、テキスト１７０が格納される。FDD１０４を介してフレキシブルディスク１０８に格納されている情報が、主メモリ１０５あるいは磁気ディスク装置１０３へ読み込まれる。

【００１０】

主メモリ１０５には、システム制御プログラム１１０、登録制御プログラム１１１、検索制御プログラム１１２、文書ファイル取得プログラム１２０、テキスト登録プログラム１２１、種文書解析プログラム１３０、テキスト読込プログラム１３１、類似度算出プログラム１３２、詳細度算出制御プログラム１３３、ブロック分割プログラム１４０、ブロック別類似度算出プログラム１４１、詳細度算出プログラム１４２、結果出力プログラム１３４及び共有ライブラリ１５０が記憶され、またワークエリア１６０が確保される。なお、共有ライブラリ１５０は、特徴語抽出プログラム１５１で構成される。

【００１１】

システム制御プログラム１１０は、登録制御プログラム１１１および検索制御プログラム１１２で構成される。登録制御プログラム１１１は、文書ファイル取得プログラム１２０およびテキスト登録プログラム１２１で構成される。検索制御プログラム１１２は、種文書解析プログラム１３０、テキスト読込プログラム１３１、類似度算出プログラム１３２、詳細度算出制御プログラム１３３および結果出力プログラム１３４で構成されるとともに、特徴語抽出プログラム１５１を呼び出す構成をとる。詳細度算出制御プログラム１３３は、ブロック分割プログラム１４０、ブロック別類似度算出プログラム１４１および詳細度算出プログラム１４２で構成されるとともに、特徴語抽出プログラム１５１を呼び出す構成をとる。

【００１２】

登録制御プログラム１１１および検索制御プログラム１１２は、ユーザによるキーボード１０１からの入力に応じてシステム制御プログラム１１０によって起動される。登録制御プログラム１１１は、文書ファイル取得プログラム１２０とテキスト登録プログラム１２１を制御する。検索制御プログラム１１２は、種文書解析プログラム１３０、特徴語抽出プログラム１５１、テキスト読込プログラム１３１、類似度算出プログラム１３２、詳細度算出制御プログラム１３３および結果出力プログラム１３４を制御する。

【００１３】

本実施例では、キーボード１０１から入力されたコマンドにより登録制御プログラム１１１および検索制御プログラム１１２が起動されるものとしたが、他の入力装置を介して入力されたコマンドあるいはイベントにより起動されるものであってもかまわない。

【００１４】

また、これらのプログラムを磁気ディスク１０３、フレキシブルディスク１０８、MO、CD-ROM、DVD等の記憶媒体（図１には示していない）に格納し、駆動装置を介して主メモリ１０５に読み込み、ＣＰＵ１０２によって実行することが可能である。また、これらのプログラムをネットワーク１０７を介して主メモリ１０５に読み込み、ＣＰＵ１０２によって実行することも可能である。

【００１５】

また、本実施例ではテキスト１７０は磁気ディスク装置１０３に格納されるものとしたが、フレキシブルディスク１０８、MO、CD-ROM、DVD等の記憶媒体（図１には示していない）に格納し、駆動装置を介して主メモリ１０５に読み込み利用することも可能であるし、あるいはネットワーク１０７を介して、他のシステムに接続された記憶媒体（図１には示していない）に格納されるものとしてもよい。また、さらにはネットワーク１０７に直接接続された記憶媒体に格納されるものとしても構わない。

【００１６】

次に、システム制御プログラム１１０の処理手順について説明する。システム制御プログラム１１０は、まずキーボード１０１から入力されたコマンドを解析する。この結果が登録実行のコマンドであると解析された場合には、登録制御プログラム１１１を起動して、文書の登録を行う。また、検索実行のコマンドであると解析された場合には、検索制御プログラム１１２を起動して、検索条件として入力された複数の単語や文、文章あるいは文書（以下、まとめて種文書と呼ぶ）に関連した内容を含む文書の検索を行う。

【００１７】

次に、システム制御プログラム１１０により起動される登録制御プログラム１１１の処理手順について説明する。登録制御プログラム１１１は、まず文書ファ

イル取得プログラム１２０を起動し、FDD１０４を介してフレキシブルディスク１０８に格納されている文書ファイルを読み込む。次に、テキスト登録プログラム１２１を起動して、前記文書ファイル取得プログラム１２０で読み込まれた文書ファイルからテキストを抽出し、磁気ディスク装置１０３にテキスト１７０として格納する。

【００１８】

なお、文書ファイルはフレキシブルディスク１０８に格納されているものとしたが、MO、CD-ROM、DVD等の記憶媒体（図１には示していない）に格納されるものとしてもよいし、ネットワーク１０７を介して、他のシステムに接続された記憶媒体（図１には示していない）に格納されるものとしてもよい。また、文書ファイル取得プログラム１２０で読み込まれた文書ファイルはテキストが抽出できるものならばよく、テキストファイルとして保存されているものであってもよいし、アプリケーションソフトの保存形式であってもよい。

【００１９】

システム制御プログラム１１０により起動される検索制御プログラム１１２の処理手順について図２を用いて説明する。検索制御プログラム１１２は、まず種文書解析プログラム１３０を起動し、検索条件で指定された種文書を読み込み、ワークエリア１６０に格納する（ステップ２００）。次に、特徴語抽出プログラム１５１を起動し、前記種文書解析プログラム１３０によりワークエリア１６０に格納された種文書から自立した意味を持つ文字列（以下、特徴語と呼ぶ）を抽出し、ワークエリア１６０に格納する（ステップ２１０）。

【００２０】

テキスト１７０に含まれるすべてのテキストに対して、ステップ２２１～ステップ２２３を繰り返し実行する（ステップ２２０）。まず、テキスト読込プログラム１３１を起動し、磁気ディスク装置１０３に格納されているテキスト１７０からテキストを１つ読み込む（ステップ２２１）。次に、類似度算出プログラム１３２を起動し、前記テキスト読込プログラム１３１により読み込まれたテキストに対し、一般的な類似文書検索技術を用いて種文書に対するテキストの類似度を算出し、ワークエリア１６０に格納する（ステップ２２２）。次に、詳細度算

出制御プログラム１３３を起動し、前記テキスト読込プログラム１３１により読み込まれたテキスト全体に対し、種文書に関する内容が占める割合（以下、詳細度と呼ぶ）を算出し、ワークエリア１６０に格納する（ステップ２２３）。

【００２１】

そして、結果出力プログラム１３４を起動し、前記類似度算出プログラム１３２により算出された類似度と前記詳細度算出制御プログラム１３３により算出された詳細度を各テキストに対して出力する（ステップ２３０）。

【００２２】

なお、特徴語抽出プログラム１５１により抽出される特徴語は、漢字やカタカナといった文字種間や文章中に存在するスペースなどの区切り文字により分割された文字列であってもよいし、形態素解析により抽出される単語やn-gramとして抽出される文字列であってもよいし、その他の方法により抽出された文字列であってもかまわない。

【００２３】

ステップ２２２における類似度算出処理は、上記従来技術に記載した類似度算出方法や、ベクトル空間法における余弦尺度を用いた類似度算出方法などを適用することができる。

【００２４】

また、類似度および詳細度が算出されるテキスト１７０は、磁気ディスク装置１０３に格納されるものとしたが、フレキシブルディスク１０８、ＭＯ、ＣＤ－ＲＯＭ、ＤＶＤ等の記憶媒体（図１には示していない）に格納されるものとして、あるいはネットワーク１０７を介して、他のシステムに接続された記憶媒体（図１には示していない）に格納されるものとしてもよい。

【００２５】

前記ステップ２２０ではテキスト１７０に含まれるすべてのテキストに対して、ステップ２２１～ステップ２２５を繰り返すものとしたが、テキスト１７０に含まれる一部のテキストに対して繰り返されるものであってもよい。

【００２６】

本実施例ではテキスト読込プログラム１３１によって読み込まれたテキスト全

体に対して類似度および詳細度を算出するものとしたが、テキスト全体でなくてもよく、テキストの一部を対象に本発明を適用することが可能である。

【００２７】

次に、検索制御プログラム１１２により起動される詳細度算出制御プログラム１３３の処理手順（図２のステップ２２３の詳細）について、図３に示すＰＡＤ図を用いて説明する。

【００２８】

まず、種文書に適合しているブロックの数（以下、適合ブロック数と呼ぶ）とテキストに含まれるブロックの数（以下、総ブロック数と呼ぶ）の初期値をともに０と設定する（ステップ３００）。次に、ブロック分割プログラム１４０を起動し、前記テキスト読込プログラム１３１で読み込まれたテキストを文、段落、章などの部分（以下、これらをまとめてブロックと呼ぶ）に分割する（ステップ３１０）。

【００２９】

前記ステップ３１０で分割された各ブロックに対して、それぞれステップ３２１～ステップ３２５を繰り返し実行する（ステップ３２０）。まず、特徴語抽出プログラム１５１を起動し、ステップ３１０で分割された各ブロックから特徴語を抽出する（ステップ３２１）。次に、ブロック別類似度算出プログラム１４１を起動し、図２のステップ２１０で抽出された種文書の特徴語と、ステップ３２１で抽出された各ブロックの特徴語から、種文書に対する各ブロックの類似度を式１を用いて算出する（ステップ３２２）。

【００３０】

【式１】

次に、ステップ３２２で算出されたブロックの類似度を、種文書に対する適合性を判定する際の基準値（以下、種文書適合性判定閾値と呼ぶ）と比較する（ステップ３２３）。この結果、ブロックの類似度が種文書適合性判定閾値以上であった場合、そのブロックを種文書に適合しているブロック（以下、適合ブロック）と判定し、適合ブロック数を１加算するとともに（ステップ３２４）、総ブロック数を１加算する（ステップ３２５）。ステップ３２３において、ブロックの

類似度が閾値以下であった場合は、適合ブロック数は１加算されず、総ブロック数のみが１加算される（ステップ３２５）。

【００３１】

ステップ３１０で分割されたすべてのブロックに対して、ステップ３２１～３２５の処理を終了したら、詳細度算出プログラム１４２を起動し、ステップ３２４およびステップ３２５で計数された適合ブロック数と総ブロック数から、式２を用いて種文書に対する該テキストの詳細度を算出する（ステップ３３０）。

【００３２】

【式２】

最後に、ステップ３３０で算出された種文書に対する該テキストの詳細度をワークエリア１６０に格納する（ステップ３４０）。

【００３３】

なお、上記ステップ３２２におけるブロックの類似度の算出には、式１に示した類似度算出式を適用したが、ベクトル空間法における余弦尺度など他の類似度算出式を適用してもよい。

【００３４】

次に、本実施例における文書検索システムの検索処理の流れについて、図４および図５を用いて説明する。

【００３５】

図４に示した例は、文書１「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」および文書２「Country-A is still in the state of economic depression. If there is bright news that induces an economic big effect, can Country-A escape from economic depression? The Sports Championship Cup was held for the first time in Countr

y-A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place on the other day. However, it was not able to become an explosive to economic recovery and an economic big effect could not be acquired.」（文書２は図４に示していない）が磁気ディスク装置１０３に格納された類似文書検索システムにおいて、種文書として「The Sports Championship Cup held for the first time in Country-A , and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place .」が入力された場合の例を示している。なお、本図は、種文書解析プログラム１３０により、検索条件として入力された種文書が文書４００として読み込まれ、テキスト読込プログラム１３１により、文書１がテキスト４１０として読み込まれた状態である。

【００３６】

まず、類似度算出プログラム１３２が実行され、前記テキスト読込プログラム１３１により読み込まれたテキスト４１０と前記種文書解析プログラム１３０により読み込まれた種文書４００から、種文書に対するテキスト４１０の類似度を算出する（図２のステップ２２２）。本実施例では、類似度を上記従来技術に記載された技術を適用して算出し、類似度算出結果４２０として類似度が“1.06”と算出され、ワークエリア１６０に格納される。ここで、種文書に含まれる文の重みはすべて“1”とする。

【００３７】

次に、ブロック分割プログラム１４０が実行され、テキスト４１０をブロック単位へ分割する（図３のステップ３１０）。本図に示した例では、テキスト４１０に対し“.”（ピリオド）を区切り文字としてブロック単位に分割しており、この結果としてブロック分割結果４３０が出力されている。本図に示したブロック分割結果４３０は、ブロック１「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」、ブロック２「Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw.」、ブロック３「Then, both the Country-C game and the Country-D game gai

ned a victory with offensive strategy, and passed the brilliant H group by the 1st place.」およびブロック４「A final tournament is due to play a match against Country-E.」であり、これらのブロックがワークエリア１６０に格納されている。

【００３８】

一方、特徴語抽出プログラム１５１が実行され、前記種文書解析プログラム１３０により読み込まれた種文書４００から“Sports”、“Championship”、“Cup”、“held”、“first”、“time”、“Country-A”、“passed”、“group”、“including”、“Country-B”、“Country-C”、“Country-D”、“1st”、“place”を特徴語４０１として抽出する（図２のステップ２１０）。また、ブロック分割結果４３０のブロック１から、“Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time”が特徴語４４０として抽出される（図３のステップ３２１）。同様に、ブロック２から“Country-A”、“played”、“match”、“against”、“first”、“game”、“Country-B”、“Championship”、“ranking”、“highest”、“group”、“though”、“troubled”、“draw”が特徴語４４１として抽出され、ブロック３からは“Country-C”、“game”、“Country-D”、“gained”、“victory”、“offensive”、“strategy”、“passed”、“brilliant”、“group”、“1st”、“place”が特徴語４４２として抽出され、ブロック４からは“final”、“tournament”、“play”、“match”、“against”、“Country-E”が特徴語４４３として抽出される。

【００３９】

次に、ブロック別類似度算出プログラム１４１が実行され、ブロック１の特徴語４４０と種文書の特徴語４０１から、種文書に対するブロック１の類似度を算出する（図３のステップ３２２）。本図で示した例では、前記特徴語抽出プログラム１５１で抽出された種文書の特徴語４０１とブロック１の特徴語４４０に関して、“Sports”、“Championship”、“Cup”、“Country-A”、“first”、“time”の６つの共通の特徴語が存在し、種文書に含まれる特徴語の個数が１５個であることから、前述の式１により、“0.40”がブロック１の類似度算出結果

４５０として算出される。

【００４０】

同様に、ブロック２～ブロック４についても、それぞれ特徴語抽出プログラム１５１で抽出された各ブロックの特徴語４４１～４４３と種文書の特徴語４０１から、ブロック別類似度算出プログラム１４１により種文書に対する各ブロックの類似度“0.33”、“0.33”、“0.00”が類似度算出結果４５１～４５３として算出される。

【００４１】

次に、上記のブロック１の類似度算出結果４５０が、あらかじめ設定された種文書適合性判定閾値以上であるか否かを判断し（図３のステップ３２３）、閾値以上であった場合、ブロック１は種文書に対する適合ブロックと判定し、適合ブロック数を１加算する（図３のステップ３２４）。本図に示した例では、種文書適合性判定閾値を“0.30”と設定しているためブロック１は適合ブロックと判定され、適合ブロック数と総ブロック数をそれぞれ１加算する（図３のステップ３２４、３２５）。

【００４２】

同様に、ブロック２～ブロック４についても、図３のステップ３２３を実行し、ブロック２とブロック３については適合ブロックと判定され、適合ブロック数と総ブロック数が１加算される。またブロック４については非適合ブロックと判定されるため、適合ブロック数は１加算せず総ブロック数のみ１加算される。

【００４３】

このように、ブロック１から順に図３のステップ３２３に示す適合ブロック判定処理を実行した後、適合ブロック数および総ブロック数の算出結果４６０～４６３が順に算出され、適合ブロック数および総ブロック数の算出結果４６３から文書１の適合ブロック数“３”および総ブロック数“４”が算出される。

【００４４】

次に、詳細度算出プログラム１４２が実行され、適合ブロック数および総ブロック数の算出結果４６３から、前述の式２を用いることにより、文書１の種文書に対する詳細度が“0.75”と算出され（図３のステップ３３０）、詳細度算出結

果４７０としてワークエリア１６０に格納される（図３のステップ３４０）。

【００４５】

同様に文書２に対しても、類似度および詳細度がそれぞれ“1.14”、“0.25”と算出される。

【００４６】

磁気ディスク装置１０３に格納されている文書１および文書２の類似度と詳細度が算出された後、結果出力プログラム１３４（図４には示していない）が実行され、ワークエリア１６０に格納されている類似度算出結果と詳細度算出結果が、検索結果一覧表示５００（図５）として出力される。図５では、結果出力として、文書１および文書２に対して文書ID、類似度、詳細度および見出しが出力されており、文書１の類似度および詳細度はそれぞれ“1.06”、“0.75”であり、文書２の類似度および詳細度はそれぞれ“1.14”、“0.25”である。

【００４７】

ここで、類似度のみでは、文書１の類似度“1.06”と文書２の類似度“1.14”であるから文書２の方を有効であると判断してしまうが、文書１の詳細度“0.75”と文書２の詳細度“0.25”から文書１の方が文書２より種文書に関する内容について全体で適合しているものと判断できる。したがって、出力された詳細度から文書１を優先して参照することで効率のいい検索が実現できる。

【００４８】

なお、図５に示した例では、検索結果一覧表示として文書ID、類似度、詳細度および見出しを出力するものとしたが、登録処理時に日付など各文書の属性情報も登録しておき、結果出力プログラム１３４でそれらの情報を出力してもよい。また、類似度および詳細度をともに出力するものとしたが、詳細度だけを出力するものとしてもよい。

【００４９】

また、図５に示した例では、各文書の出力順は類似度の降順で出力するものとしたが、詳細度の降順で出力するものとしてもよいし、これらを図６に示すように表示オプションで選択できるようにしてもよい。図６に示した例では、表示オプションとして類似度の降順で表示するかあるいは詳細度の降順で表示するかを

選択可能としたインターフェースを備えており、図６では詳細度順が選択されていることにより詳細度の高い順に文書１と文書２が表示されている。

【００５０】

また、図５および図６に示した例では、テキスト１７０として磁気ディスク１０３に格納されているすべてのテキストに対して結果を表示するものとしたが、図７に示すように検索者およびシステム管理者によって予め設定された類似度および詳細度に関する閾値により、検索結果として表示する対象文書を決定してもよい。図７に示した例では、類似度および詳細度に関する閾値を設定するインターフェースを備えており、類似度の閾値が“0.00”以上および詳細度の閾値が“0.50”以上と設定されているため、その条件を満たしている文書１のみの結果が表示されている。

【００５１】

また、図５、図６および図７に示した例では、類似度および詳細度が検索結果の一覧表示で出力されるものとしたが、図８のように指定された文書の全文が表示されるとととともに、類似度あるいは詳細度の少なくとも一方が出力されるようにしてもよい。図８では、文書１の全文を表示するとともに、類似度および詳細度を表示して出力している。また、類似度および詳細度に関してあらかじめ設定された閾値以上の文書に対しては図８に示すように類似度、詳細度および全文が出力され、閾値以下の文書に対しては図５および図６に示すように一覧表示として文書ID、類似度、詳細度、見出しが出力されるものとしてもよい。

【００５２】

また、種文書に対する対象文書の類似度算出方法において、類似度算出プログラム１３２を実行せずに（つまり図２のステップ２２２を実行せずに）、ブロック別類似度算出プログラム１４１で算出された類似度結果４５０～４５３を加算することにより（図３のステップ３２２）、対象文書の類似度を算出してもよい。

【００５３】

また、本実施例ではテキスト１７０のすべてのテキストに対して、類似度算出プログラム１３２（図２のステップ２２２）および詳細度算出プログラム１３３

（同ステップ２２３）を実行したが、類似度算出プログラム１３２で算出された類似度があらかじめ設定された閾値以上のテキストに対して詳細度算出プログラム１３３を実行してもよい。逆に、詳細度算出プログラム１３３で算出された詳細度があらかじめ設定された閾値以上のテキストに対して類似度算出プログラム１３２を実行してもよい。これにより、類似度あるいは詳細度の算出対象となるテキスト数を削減することができ、高速に検索を行うことができる。

【００５４】

また、本実施例では、予め蓄積された文書に対して検索条件との関連性を判定する文書検索システムとして説明したが、特開平２０００－３３９３４６号公報に記載されている類似文書検索配送システムにおける適合度算出プログラムを、本発明における詳細度算出制御プログラムに置き換えてもよい。

【００５５】

このように本発明による詳細度は、予め蓄積された文書に対して検索条件との関連性を判定する文書検索システムだけでなく、１件の対象文書に対して配信条件との関連性を判定する文書配信システムにも適用できる。

【００５６】

以上説明したように、本発明の第一の実施例によれば、種文書に関する内容について、対象文書の全体で類似しているのか、あるいは対象文書の一部で類似しているのかを判断できるため、有効な文書を効率よく検索できるようになる。

【００５７】

次に、本発明の第二の実施例について説明する。第二の実施例では、検索条件として種文書と全文検索条件の両方が指定された場合における詳細度を算出する。

【００５８】

本実施例の文書検索システムは、図１に示した第一の実施例のシステムとほぼ同様の構成であるが、検索制御プログラム１１２と詳細度算出プログラム１３３が異なり、図９に示すように、検索制御プログラム１１２ｃには全文検索条件解析プログラム１３０ａが加わるとともに、詳細度算出プログラム１３３０にはブロック別全文検索条件適合度算出プログラム１４１ａが加わる。

【００５９】

以下、第一の実施例と異なる検索制御プログラム１１２ｃの処理手順について図１０を用いて説明する。ここで第一の実施例（図２）と異なるのは、種文書解析プログラム１３０が実行された後に全文検索条件解析プログラム１３０ａが実行されること、及び類似度算出プログラム１３２が実行された後に詳細度算出制御プログラム１３３０が実行されることである。

【００６０】

検索制御プログラム１１２ｃは、まず種文書解析プログラム１３０を起動し、検索条件で指定された種文書を読み込み、ワークエリア１６０に格納する（ステップ２００）。次に、全文検索条件解析プログラム１３０ａを起動し、検索条件で指定された全文検索条件を読み込む。この全文検索条件に含まれるAND、OR、NOTの論理演算子を識別することによりその構造を解析し、和積標準形で表された論理演算式（以下、解析済論理演算式と呼ぶ）をワークエリア１６０に格納する（ステップ２００ａ）。次に、特徴語抽出プログラム１５１を起動し、前記種文書解析プログラム１３０によりワークエリア１６０に格納された種文書から特徴語を抽出し、ワークエリア１６０に格納する（ステップ２１０）。

【００６１】

次に、テキスト１７０に含まれるすべてのテキストに対して、ステップ２２１～ステップ２２３を繰り返し実行する（ステップ２２０）。まず、テキスト読込プログラム１３１を起動し、磁気ディスク装置１０３に格納されているテキスト１７０からテキストを１つ読み込む（ステップ２２１）。次に、類似度算出プログラム１３２を起動し、前記テキスト読込プログラム１３１により読み込まれたテキストに対し、種文書に対するテキストの類似度を算出し、ワークエリア１６０に格納する（ステップ２２２）。詳細度算出制御プログラム１３３０を起動し、検索条件に対する前記テキスト読込プログラム１３１により読み込まれたテキストの詳細度を算出し、ワークエリア１６０に格納する（ステップ２２３ｃ）。

【００６２】

そして、結果出力プログラム１３４を起動し、前記類似度算出プログラム１３２により算出された類似度と前記詳細度算出制御プログラム１３３０により算出

された詳細度を各テキストに対して出力する（ステップ２３０）。

【００６３】

次に、詳細度算出制御プログラム１３３０の処理手順について図１１を用いて説明する。ここで第一の実施例（図３）と異なるのは、ブロック別類似度算出プログラム１４１が実行された後にブロック別全文検索条件適合度算出プログラム１４１ａが実行されることと、図３に示す適合性判定ステップ３２３において、種文書適合性判定閾値のみを適合ブロック判定基準に用いるのではなく、ブロック別全文検索条件適合度算出プログラム１４１ａによって算出された全文検索条件適合度に関する閾値（以下、全文検索条件適合性判定閾値と呼ぶ）も適合ブロック判定基準に用いることである。

【００６４】

まず、テキストの適合ブロック数とテキストに含まれる総ブロック数の初期値をともに０に設定する（ステップ３００）。ブロック分割プログラム１４０を起動し、ステップ２２１（図１０）において読み込まれたテキストをブロックに分割する（ステップ３１０）。

【００６５】

次に、ステップ３１０で分割された各ブロックに対して、それぞれステップ３２１～３２５を繰り返し実行する（ステップ３２０）。まず、特徴語抽出プログラム１５１を起動し、各ブロックから特徴語を抽出する（ステップ３２１）。次に、ブロック別類似度算出プログラム１４１を起動し、特徴語抽出プログラム１５１により抽出された種文書の特徴語と前記ステップ３２１で抽出された各ブロックの特徴語から、種文書に対するブロックの類似度を式１を用いて算出する（ステップ３２２）。

【００６６】

【式１】

式1

$$類似度＝\frac{種文書とブロックの共通の特徴語の数}{種文書の特徴語の数}$$

次に、ブロック別全文検索条件適合度算出プログラム１４１ａを起動し、全文検索条件解析プログラム１３０ａにより読み込まれた解析済論理演算式から、全文検索条件に対するブロックの適合度（以下、全文検索条件適合度と呼ぶ）を算出する（ステップ３２２ａ）。

【００６７】

次に、前記ブロック別類似度算出プログラム１４１により算出された各ブロックの類似度を、種文書適合性判定閾値と比較するとともに、ステップ３２２ａで算出されたブロックの全文検索条件適合度を、全文検索条件適合性判定閾値と比較する（ステップ３２３ｃ）。この比較の結果、あるブロックの類似度が種文書適合性判定閾値以上であり、かつそのブロックの全文検索条件適合度が全文検索条件適合性判定閾値以上の場合、そのブロックを検索条件に対する適合ブロックと判定し、適合ブロック数を１加算するとともに（ステップ３２４）、総ブロック数を１加算する（ステップ３２５）。ステップ３２３ｃにおいて適合度または類似度のどちらかが閾値以下であった場合は、適合ブロック数は１加算されず、総ブロック数のみが１加算される（ステップ３２５）。

【００６８】

次に、詳細度算出プログラム１４２を起動し、前記ステップ３２４およびステップ３２５で計数された適合ブロック数と総ブロック数から、式２を用いて種文書に対する該テキストの詳細度を算出する（ステップ３３０）。

【００６９】

【式２】

式2

$$詳細度＝\frac{適合ブロック数}{総ブロック数}$$

最後に、前記ステップ３３０で算出された種文書に対する該テキストの詳細度をワークエリア１６０に格納する（ステップ３４０）。

【００７０】

次に、詳細度算出制御プログラム１３３０により起動されるブロック別全文検索条件適合度算出プログラム１４１ａの処理手順について説明する。まず、全文検索条件解析プログラム１３０ａにより和積標準形でワークエリア１６０に読み込まれた解析済論理演算式に対し、AND演算子を境界として分割される単語や論理演算式（以下、部分論理演算式）を抽出する。次に、特徴語抽出プログラム１５１により抽出された処理対象となるブロックの特徴語が、抽出された各部分論理式の条件と適合するかどうかを判定する。

【００７１】

この結果、処理対象のブロックが満たす部分論理演算式の数（以下、適合部分論理式数と呼ぶ）と、解析済論理演算式に含まれる部分論理演算式（以下、総部分論理演算式数と呼ぶ）を計数し、式３より全文検索条件に対するブロックの全文検索条件適合度を算出する。

【００７２】

【式３】

式3

$$\text{全文検索条件適合度}=\frac{\text{適合部分論理演算式数}}{\text{総部分論理演算式数}}$$

なお、ステップ３２２ａにおける、ブロック別全文検索条件適合度算出プログラム１４１ａによるブロックの全文検索条件適合度の算出には、指定された全文検索条件に含まれる部分論理演算式の総数に対し、該ブロックの特徴語により満たされている部分論理式の数の割合を算出したが、特開平１１－１５４１６４号公報や特開２００１－８４２５５号公報に開示されている方法を用いてもよい。

【００７３】

以下、本実施例の検索処理におけるブロックの適合性判定について、具体的な処理の流れを図１２を用いて説明する。

【００７４】

本図に示した例は、文書１「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」が磁気ディスク装置１０３に格納された文書検索システムにおいて、種文書として「The Sports Championship Cup held for the first time in Country-A , and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place.」、全文検索条件として「“Country-A” and “Country-B” and (“Championship” or “tournament”)」が入力された場合の例を示している。なお、本図は、種文書解析プログラム１３０により検索条件として入力された種文書が文書４００と

して読み込まれ、全文検索条件解析プログラム１３０ａにより検索条件として入力された全文検索条件が解析済論理演算式４０００として読み込まれ、テキスト読込プログラム１３１により文書１がテキスト４１０として読み込まれた状態である。

【００７５】

まず、特徴語抽出プログラム１５１が実行され、前記種文書解析プログラム１３０により読み込まれた種文書４００から、“Sports”、“Championship”、“Cup”、“held”、“first”、“time”、“Country-A”、“passed”、“group”、“including”、“Country-B”、“Country-C”、“Country-D”、“1st”、“place”を特徴語４０１として抽出する（図１０のステップ２１０）。次に、ブロック分割プログラム１４０が実行され、テキスト４１０をブロック単位へ分割する（図１１のステップ３１０）。本図に示した例では、テキスト４１０を“.”（ピリオド）を区切り文字としてブロック単位に分割しており、この分割結果から「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」がブロック１の抽出結果４３００として出力されている。

【００７６】

次に、特徴語抽出プログラム１５１が実行され、ブロック分割プログラム１４０で文書１より分割されたブロック１から、“Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time”を特徴語４４０として抽出する（図１１のステップ３２１）。次に、ブロック別類似度算出プログラム１４１が実行され、ブロック１の特徴語４４０と種文書の特徴語４０１から、種文書に対するブロック１の類似度を算出する（図１１のステップ３２２）。本図で示した例では、特徴語抽出プログラム１５１で抽出された種文書の特徴語４０１とブロック１の特徴語４４０の間で、“Sports”、“Championship”、“Cup”、“Country-A”、“first”、“time”の６つの共通の特徴語が存在し、種文書に含まれる特徴語の個数が１５個であることから、前述した式１より、“0.40”がブロック１の類似度算出結果４５０として算出される。

【００７７】

次に、ブロック別全文検索条件適合度算出プログラム１４１aが実行され、全文検索条件に対するブロック１の全文検索条件適合度を算出する（図１１の３２２ａ）。本図で示した例では、ブロック１の特徴語４４０には“Country-A”および“Championship”が含まれており、解析済論理演算式４０００「“Country-A” and “Country-B” and（“Championship” or “tournament”）」の部分論理演算式「“Country-A”」、「“Championship” or “tournament”」を満たしている。すなわち、解析済論理演算式4000に含まれる３つの部分論理演算式のうち、２つが満たされていることから、“0.67”がブロック１の全文検索条件適合度算出結果４５００として算出される。

【００７８】

そして、ブロック１の類似度が種文書適合性判定閾値以上であり、かつブロック１の全文検索条件適合度が全文検索条件適合性閾値以上であるかどうかを判定する（図１１のステップ３２３ｃ）。判定の結果、両方の値が閾値以上である場合は、ブロック１は検索条件に対して適合ブロックと判定される。本図に示した例では、種文書適合性閾値および全文検索条件適合性閾値をそれぞれ“0.30”としており、ブロック１の類似度“0.40”、および詳細度 “0.67”はそれぞれこの条件を満たしているため適合ブロックと判定される。

【００７９】

次に、図１２に示した、ブロック別全文検索条件適合度算出プログラム１４１ａが行うブロック別全文検索条件適合度算出処理（図１１のステップ３２２ａ）の詳細について、図１３を用いて説明する。

【００８０】

本図に示した例では、全文検索条件解析プログラム１３０ａによって読み込まれた解析済論理式４０００「“Country-A” and “Country-B” and（“Championship” or “tournament”）」に対し、図１２に示したブロック１の特徴語４４０からブロック１の全文検索適合度を算出する場合の処理の流れを示している。

【００８１】

まず、解析済論理演算式４０００から部分論理演算式４５０１を抽出する（ス

テップ３２２１）。ここでは、和積標準形で読み込まれた解析済論理演算式がAND演算子を境界として分割され、その分割された単語や論理演算式を、部分論理式として抽出する。本図に示した例では、AND演算子を境界として、解析済論理演算式４０００から「“Country-A”」、「“Country-B”」、「“Championship”or“tournament”」が抽出される。

【００８２】

次に、ブロック１の特徴語４４０と前記部分論理演算式抽出ステップ３２２１によって抽出された部分論理演算式４５０１から、各部分論理演算式に対するブロックの適合判定を行う（ステップ３２２１）。そして、判定結果４５０２を出力する。本図に示した例では、ブロック１の特徴語が“Country-A”、“Championship”を含むことから、ブロック１を満たす部分論理演算式４５０１は「“Country-A”」、「“Championship”or“tournament”」と判定される。

【００８３】

次に、解析済論理演算式４０００に対するブロック１の全文検索条件適合度４５００を算出する（ステップ３２２３）。本図に示した例では、前記部分論理演算式適合判定ステップ３２２２によるブロックの適合判定結果４５０２から、部分論理式数“3”が計数されると共に、ブロック１が満たす部分論理式数“2”と計数される。この結果、式３より“0.67”が全文検索条件適合度４５００として算出される。

【００８４】

以上説明したように本発明の第二の実施形態によれば、種文書の内容に対する類似性と全文検索条件に対する適合性の両方を用いて詳細度の算出を行うことにより、検索者の検索目的に応じた、より精度の高い検索条件に関する文書の詳細度を算出することができる。

【００８５】

なお本実施例では、検索条件として種文書と全文検索条件の両方を指定する構成を採用したが、全文検索条件のみが指定される場合でもよい。その場合、図９に示した種文書解析プログラム１３０とブロック別類似度算出プログラム１４１がなくなるとともに、図１１に示したステップ３２３ｃの適合ブロックの判定処

理に関する判定基準が全文検索条件適合度のみとなる。また、図１０に示したステップ２２２における類似度算出処理は、全文検索条件に関するテキストの類似度として、拡張ブーリアンに基づいた方法や、特開平１１－１５４１６４号公報に基づいた方法で算出される。

【００８６】

次に、第三の実施例について説明する。第三の実施例では、文書ファイルの登録時にブロックごとに抽出された特徴語を、あらかじめブロック別特徴語ファイルとして格納しておき、詳細度の算出時には、そのブロック別特徴語ファイルを読み込むことで詳細度を算出する。

【００８７】

本実施例の文書検索システムは、図１に示した第一の実施例のシステムとほぼ同様の構成を取るが、図１４に示すように磁気ディスク装置１０３にブロック別特徴語ファイル１７１が追加されるとともに、登録制御プログラム１１１と詳細度算出制御プログラム１３３の構成が異なり、登録制御プログラム１１１ｃにはブロック分割プログラム１４０とブロック別特徴語登録プログラム１２００が加わるとともに、詳細度算出制御プログラム１３３１にはブロック分割プログラム１４０の代りに特徴語読込プログラム１４００が加わる。

【００８８】

以下、第一の実施例とは異なる登録制御プログラム１１１ｃの処理手順を図１５を用いて説明する。ここで、第一の実施例と異なるのは、テキスト登録プログラム１２１が実行された後に、ブロック別特徴語ファイル１７１を作成するために、ブロック分割プログラム１４０、特徴語抽出プログラム１５１およびブロック別特徴語登録プログラム１２００が実行されることである。

【００８９】

登録制御プログラム１１１ｃでは、まず文書ファイル取得プログラム１２０を起動し、FDD１０４を介してフレキシブルディスク１０８に格納されている文書ファイルをワークエリア１６０に読み込む（ステップ７００）。次に、テキスト登録プログラム１２１を起動して、ステップ７００で読み込まれた文書ファイルからテキストを抽出し、ワークエリア１６０に格納するとともにテキスト１７０

として磁気ディスク装置１０３に格納する（ステップ７１０）。次に、ブロック分割プログラム１４０を起動し、ステップ７１０でワークエリア１６０に格納されたテキストをブロック単位に分割する（ステップ７２０）。

【００９０】

次に、ステップ７２０で分割された各ブロックに対して、それぞれステップ７３１～ステップ７３２を繰り返し行う（ステップ７３０）。まず、特徴語抽出プログラム１５１を起動し、各ブロックの特徴語を抽出する（ステップ７３１）。次に、ブロック別特徴語ファイル作成プログラム１２００を起動し、ステップ７３１により各ブロックから抽出された特徴語を、ブロック別特徴語ファイル１７１に登録する（ステップ７３２）。

【００９１】

以下、第一の実施例と異なる詳細度算出制御プログラム１３３１の処理手順を図１６を用いて説明する。第一の実施例における詳細度算出制御プログラム１３３の処理手順（図３）と異なるのは、ステップ３１０がなくなるとともに、ステップ３２１の代りにステップ３２１ａが加わることである。

【００９２】

まず、詳細度算出制御プログラム１３３１は、まず適合ブロック数と総ブロック数の初期値をともに０と設定する（ステップ３００）。次に、１つのテキストに含まれるすべてブロックに対して、それぞれステップ３２１ａ～ステップ３２５を繰り返し実行する（ステップ３２０）。

【００９３】

まず、特徴語読込プログラム１４００を起動し、ブロック別特徴語ファイル１７１から１ブロック分の特徴語を読み込む（ステップ３２１ａ）。次に、ブロック別類似度算出プログラム１４１を起動し、上述した式１より種文書に対するブロックの類似度を算出する（ステップ３２２）。次に、ステップ３２２で算出されたブロックの類似度を種文書適合性判定閾値と比較する（ステップ３２３）。この結果、ブロックの類似度が種文書適合性判定閾値以上であった場合、そのブロックは適合ブロックと判定され、適合ブロック数を１加算するとともに（ステップ３２４）、総ブロック数を１加算する（ステップ３２５）。ステップ３２３

において閾値以下であった場合は、適合ブロック数は１加算されず、総ブロック数のみが１加算される（ステップ３２５）。

【００９４】

次に、詳細度算出プログラム１４２を起動し、ステップ３２４およびステップ３２５で計数された適合ブロック数と総ブロック数から、式２を用いて種文書に対するそのテキストの詳細度を算出する（ステップ３３０）。次に、ステップ３３０で算出された種文書に対するそのテキストの詳細度をワークエリア１６０に格納する（ステップ３４０）。

【００９５】

次に、文書の登録処理におけるブロック別の特徴語をディスク装置１０３のブロック別特徴語ファイル１７１に登録する処理の流れについて、図１７を用いて説明する。本図に示した例では、文書１「In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.」および文書２「Country-A is still in the state of economic depression. If there are bright news that induce an economic big effect, can Country-A escape from economic depression? The Sports Championship Cup was held for the first time in Country-A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place on the other day. However, it was not able to become an explosive to economic recovery and an economic big effect could not be acquired.」が、テキスト読込プログラム１３１により、それぞれテキスト４１０およびテキスト９００として読み込まれた状態から、文書１および文書２の各ブロックの特徴語をブロック別特徴語ファイル１７１に登録する処理の流れを説明している。

【００９６】

まず、ブロック分割プログラム１４０が実行され、テキスト読込プログラム１３１により読み込まれたテキスト４１０をブロック単位に分割する。本図に示した例では、“.”（ピリオド）を区切り文字としてテキスト４１０をブロック単位に分割しており、この結果としてブロック分割結果４３０が出力される。図１７に示したブロック分割結果４３０は、ブロック１「In The Sports Championship Cup, Country-A broke through the primary league for the first time.」、ブロック２「Country-A played a match against the first game and Country-B of the Championship ranking highest in H group, and though troubled, and was a draw.」、ブロック３「Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the lst place.」およびブロック４「A final tournament is due to play a match against Country-E.」が格納されていることを表している。

【００９７】

次に、特徴語抽出プログラム１５１が実行され、ブロック分割結果４３０のブロック１から、特徴語４４０として“Sports”、“Championship”、“Cup”、“Country-A”、“broke”、“through”、“primary”、“league”、“first”、“time”を抽出する。そして、ブロック別特徴語登録プログラム１２００が実行され、前記特徴語抽出プログラム１５１により抽出されたブロック１の特徴語４４０は、文書１のブロック１の特徴語として、ブロック別特徴語ファイル１７１に登録される。また、合わせて文書ID“１”およびブロックID“１”もブロック別特徴語ファイル１７１に登録される。

【００９８】

同様にブロック２～ブロック４についても特徴語抽出プログラム１５１により特徴語４４１～４４３が抽出され、各ブロックにおいて抽出された特徴語がそれぞれ文書１の各ブロックの特徴語としてブロック別特徴語ファイル１７１に登録される。

【００９９】

同様に文書２についても、テキスト読込プログラム１３１で読み込まれたテキ

スト９００に対し、ブロック分割プログラム１４０によりブロック分割結果９０１が出力され、特徴語抽出プログラム１５１により各ブロックから特徴語９４０～９４３が抽出され、抽出された特徴語が、ブロック別特徴語登録プログラム１２００によりそれぞれ文書２の各ブロックの特徴語としてブロック別特徴語ファイル１７１に登録される。

【０１００】

なお、本図のブロック別特徴語ファイル１７１に格納されている文書ID“１”および“２”は、それぞれ文書１および文書２に対応している。

【０１０１】

以上説明したように、本発明の第三の実施例によれば、ブロック別特徴語ファイル１７１を文書登録時にあらかじめ作成しておくことにより、検索の度にテキストのブロック分割処理およびブロックの特徴語抽出処理を実行する必要がないため、検索時には大量のテキストに対しても高速に詳細度の算出を行うことができる。

【０１０２】

なお、本実施例においては、テキスト読込プログラム１３１を起動してテキスト１７０を読み込み、類似度を算出する構成としたが、テキスト読込プログラム１３１を呼び出さず、検索制御プログラム１１２が特徴語読込プログラム１４００を呼び出し、ブロック別特徴語ファイル１７１を読み込んだ値を用いて類似度を算出してもよい。これにより、テキストを読み込まなくてもよくなるため、メモリの使用量を軽減することができる。

【０１０３】

【発明の効果】

以上説明したように本発明によれば、種文書に関する対象文書の類似度だけでなく、対象文書全体に対して種文書の内容が占める割合を表す詳細度が出力されるようになる。これにより、種文書に関する内容について、対象文書の全体で類似しているのか、あるいは対象文書の一部で類似しているのかを容易に判断できるため、文書を効率よく検索できる。

【図面の簡単な説明】

【図１】本発明の第一の実施例における類似文書検索システムの全体構成を示す図である。

【図２】本発明の第一の実施例における検索制御プログラム１１２の処理を示すＰＡＤ図である。

【図３】本発明の第一の実施例における詳細度算出制御プログラム１３３の処理を説明するＰＡＤ図である。

【図４】本発明の第一の実施例における検索制御プログラム１１２の具体的な処理の流れを説明する図である。

【図５】本発明の第一の実施例における検索結果一覧画面を示す図である。

【図６】本発明の第一の実施例における検索結果一覧画面を示す図である。

【図７】本発明の第一の実施例における結果出力プログラム１３４の出力対象文書として、類似度および詳細度の閾値を設定する検索結果一覧画面を示す図である。

【図８】本発明の第一の実施例における、対象文書の全文を表示する画面を示す図である。

【図９】本発明の第二の実施例における類似文書検索システムの全体構成を示す図である。

【図１０】本発明の第二の実施例における検索制御プログラム１１２ｃの処理を説明するＰＡＤ図である。

【図１１】本発明の第二の実施例における詳細度算出制御プログラム１３３０の処理を説明するＰＡＤ図である。

【図１２】本発明の第二の実施例の検索制御プログラム１１２ｃにおける適合ブロック判定処理の具体的な流れを説明する図である。

【図１３】本発明の第二の実施例の全文検索条件適合度算出プログラム１４１ａの具体的な処理の流れを説明する図である。

【図１４】本発明の第三の実施例における類似文書検索システムの全体構成を示す図である。

【図１５】本発明の第三の実施例における登録制御プログラム１１１の処理を説明するＰＡＤ図である。

【図１６】本発明の第三の実施例における詳細度算出制御プログラム１３３１の処理を説明するＰＡＤ図である。

【図１７】本発明の第三の実施例における登録制御プログラム１１１の具体的な処理の流れを説明する図である。

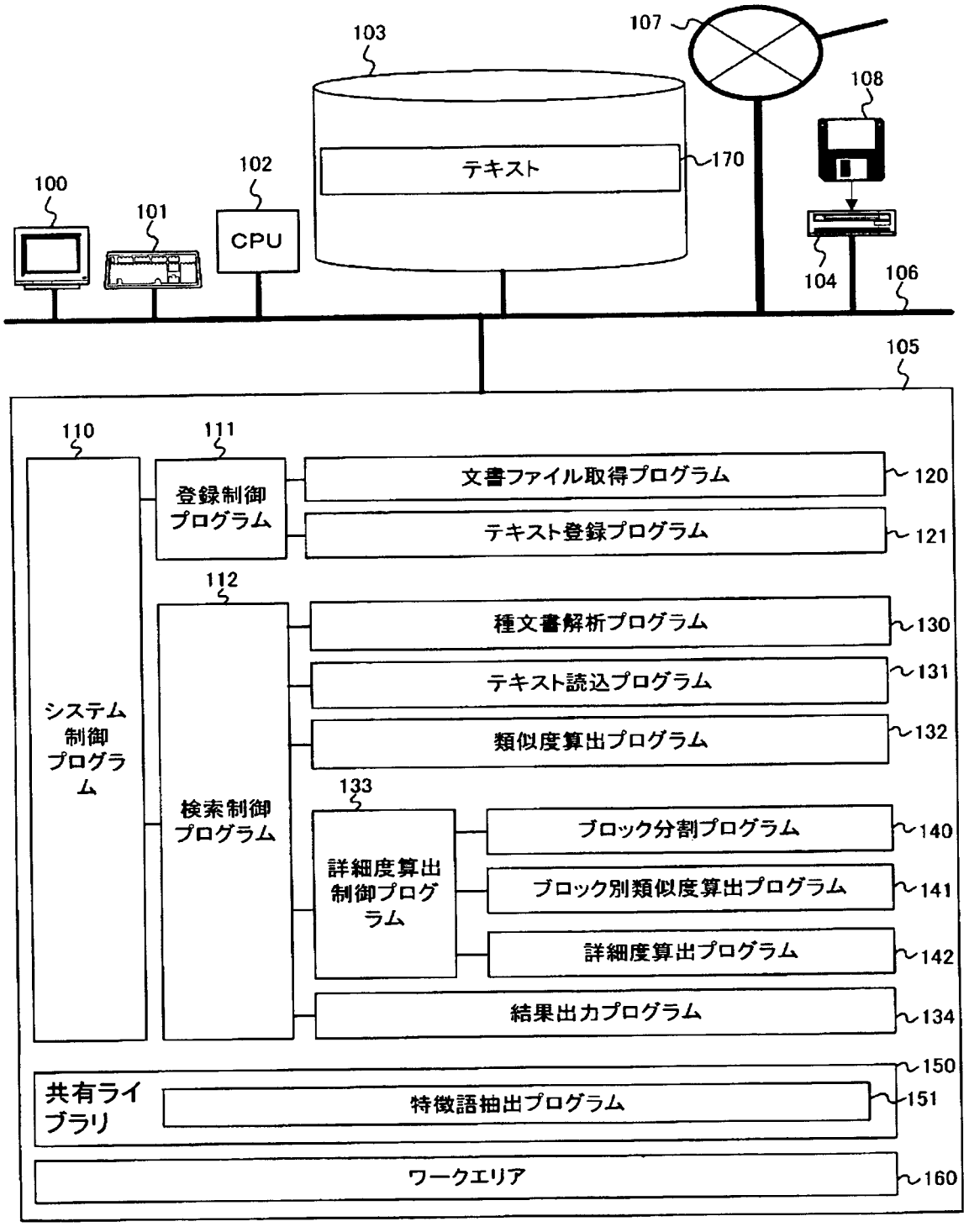
【符号の説明】

１００…ディスプレイ、１０１…キーボード、１０２…中央演算処理装置（ＣＰＵ）、１０３…磁気ディスク装置、１０４…フレキシブルディスクドライブ（ＦＤＤ）、１０５…主メモリ、１１０…システム制御プログラム、１１１…登録制御プログラム、１１２…検索制御プログラム、１２０…文書ファイル取得ファイル、１２１…テキスト登録プログラム、１３０…種文書解析プログラム、１３１…テキスト読込プログラム、１３２…類似度算出プログラム、１３３…詳細度算出制御プログラム、１３４…結果出力プログラム、１４０…ブロック分割プログラム、１４１…ブロック別類似度算出プログラム、１４２…詳細度算出プログラム、１５０…共有ライブラリ、１５１…特徴語抽出プログラム、

【書類名】　図面

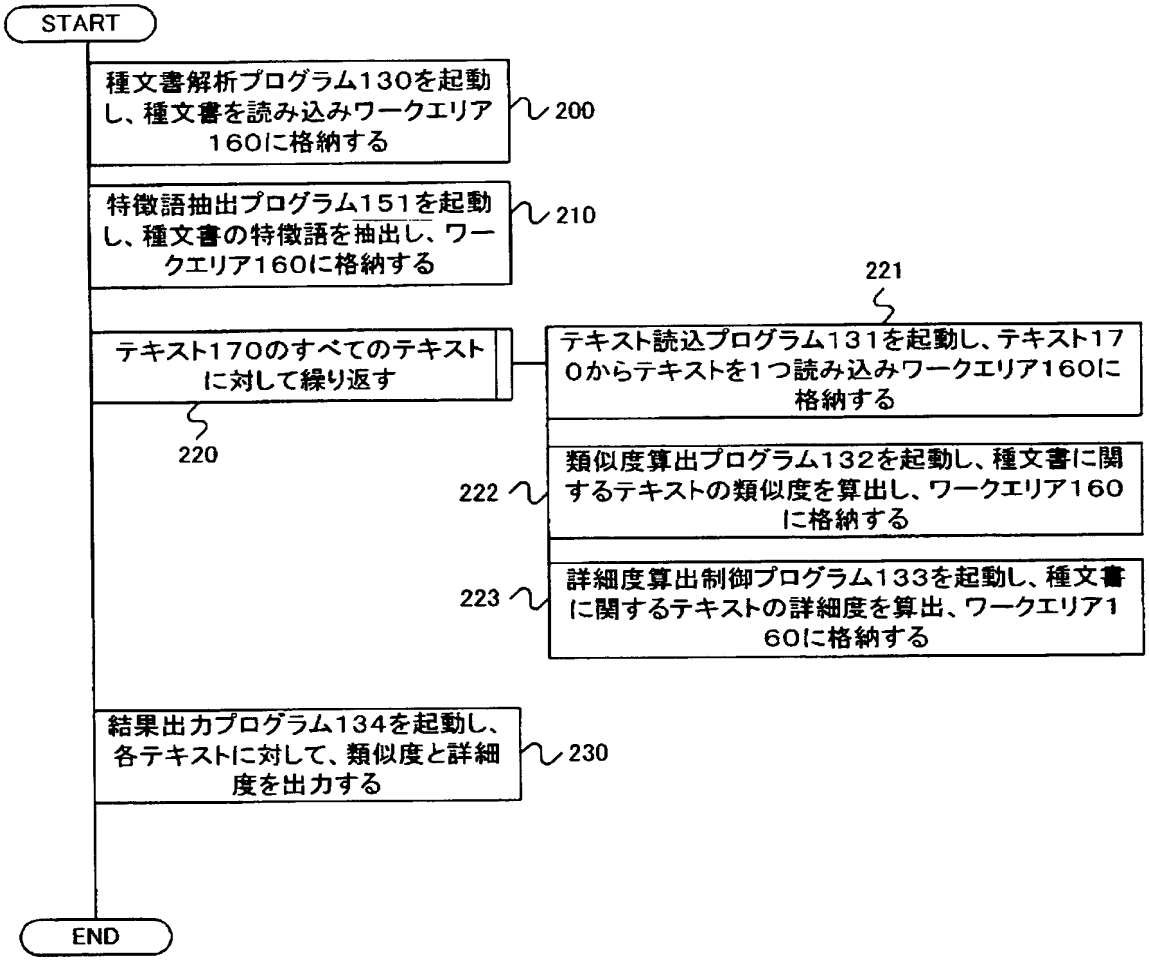
【図１】

図1


103
107
108
テキスト
170
100
101
102
CPU
104
106
105
110
111
登録制御プログラム
文書ファイル取得プログラム
120
テキスト登録プログラム
121
112
種文書解析プログラム
130
テキスト読込プログラム
131
類似度算出プログラム
132
システム制御プログラム
検索制御プログラム
133
詳細度算出制御プログラム
ブロック分割プログラム
140
ブロック別類似度算出プログラム
141
詳細度算出プログラム
142
結果出力プログラム
134
150
共有ライブラリ
特徴語抽出プログラム
151
ワークエリア
160

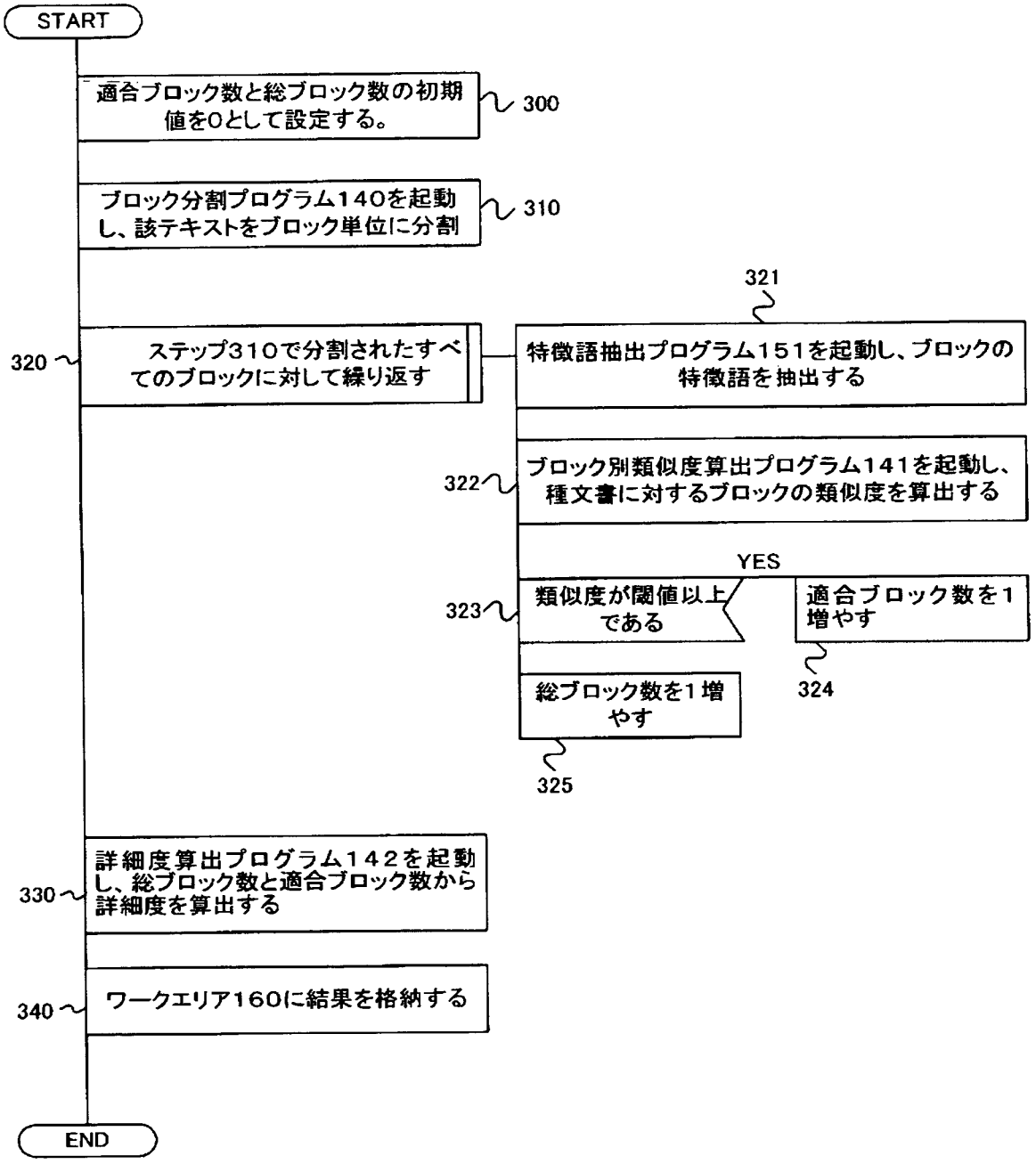
【図２】

図2


START
種文書解析プログラム130を起動し、種文書を読み込みワークエリア160に格納する
200
特徴語抽出プログラム151を起動し、種文書の特徴語を抽出し、ワークエリア160に格納する
210
221
テキスト170のすべてのテキストに対して繰り返す
220
テキスト読込プログラム131を起動し、テキスト170からテキストを1つ読み込みワークエリア160に格納する
222
類似度算出プログラム132を起動し、種文書に関するテキストの類似度を算出し、ワークエリア160に格納する
223
詳細度算出制御プログラム133を起動し、種文書に関するテキストの詳細度を算出、ワークエリア160に格納する
結果出力プログラム134を起動し、各テキストに対して、類似度と詳細度を出力する
230
END

【図３】

図3


START
適合ブロック数と総ブロック数の初期値を0として設定する。
300
ブロック分割プログラム140を起動し、該テキストをブロック単位に分割
310
321
320
ステップ310で分割されたすべてのブロックに対して繰り返す
特徴語抽出プログラム151を起動し、ブロックの特徴語を抽出する
ブロック別類似度算出プログラム141を起動し、種文書に対するブロックの類似度を算出する
322
YES
323
類似度が閾値以上である
適合ブロック数を1増やす
324
総ブロック数を1増やす
325
詳細度算出プログラム142を起動し、総ブロック数と適合ブロック数から詳細度を算出する
330
ワークエリア160に結果を格納する
340
END

【図４】

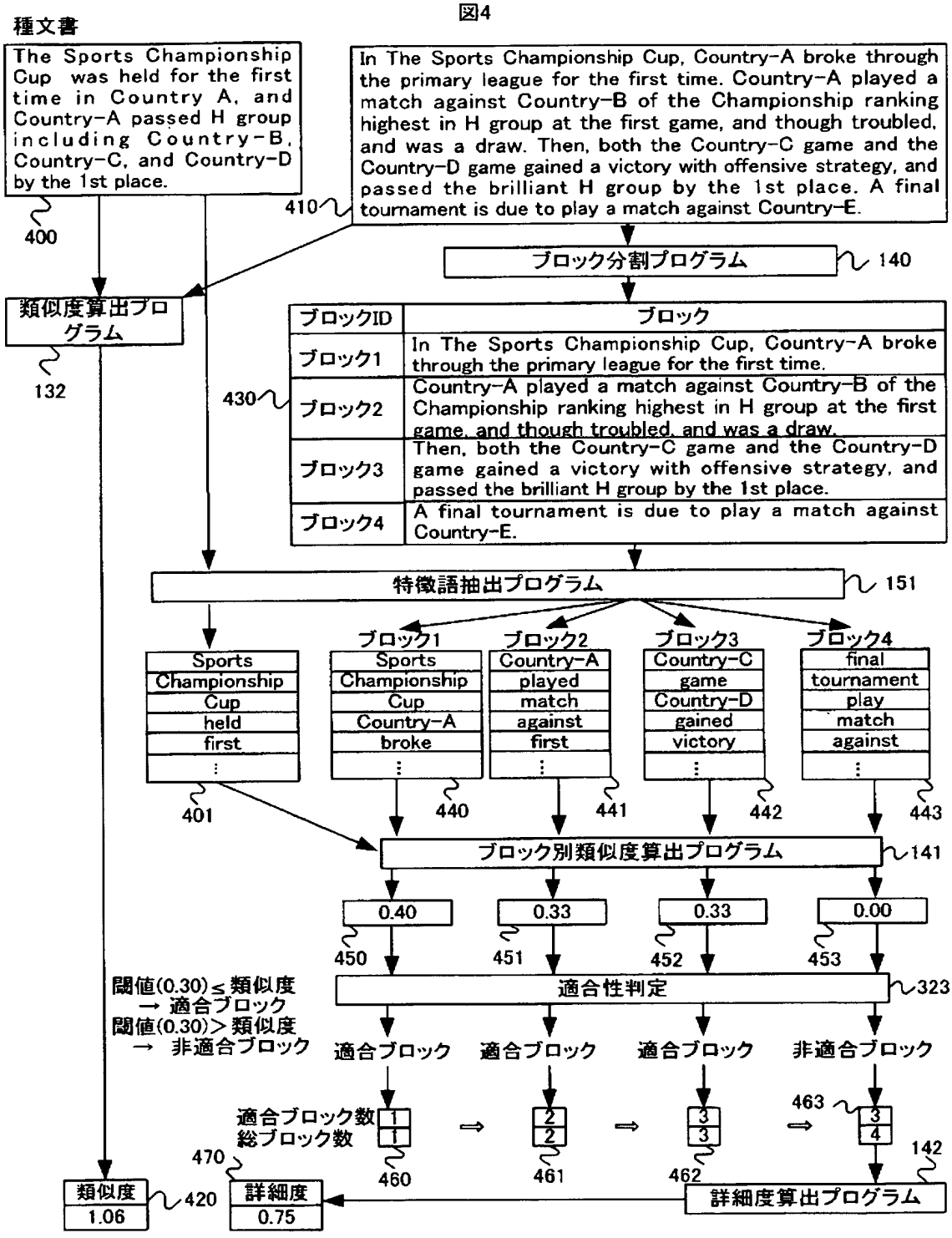


図4
種文書
The Sports Championship Cup was held for the first time in Country A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place.
400
In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.
410
ブロック分割プログラム
140
類似度算出プログラム
132
ブロックID
ブロック
ブロック1
In The Sports Championship Cup, Country-A broke through the primary league for the first time.
ブロック2
Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw.
ブロック3
Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place.
ブロック4
A final tournament is due to play a match against Country-E.
430
特徴語抽出プログラム
151
ブロック1
ブロック2
ブロック3
ブロック4
Sports
Championship
Cup
held
first
Sports
Championship
Cup
Country-A
broke
Country-A
played
match
against
first
Country-C
game
Country-D
gained
victory
final
tournament
play
match
against
401
440
441
442
443
ブロック別類似度算出プログラム
141
0.40
0.33
0.33
0.00
450
451
452
453
閾値(0.30)≦類似度
→ 適合ブロック
閾値(0.30)＞類似度
→ 非適合ブロック
適合性判定
323
適合ブロック
適合ブロック
適合ブロック
非適合ブロック
適合ブロック数
総ブロック数
1
1
2
2
3
3
3
4
463
460
461
462
142
470
類似度
1.06
420
詳細度
0.75
詳細度算出プログラム

【図５】

図5

検索結果一覧

文書ID	類似度	詳細度	見出し
2	1.14	0.25	Country-A is still in the state of economic …
1	1.06	0.75	In The Sports Championship Cup, Country-A …

500

【図６】

図6

検索結果一覧

表示順　○類似度順　◉詳細度順

501

文書ID	詳細度	類似度	見出し
1	0.75	1.06	In The Sports Championship Cup, Country-A …
2	0.25	1.14	Country-A is still in the state of economic …

【図７】

図7

検索結果一覧

表示文書条件

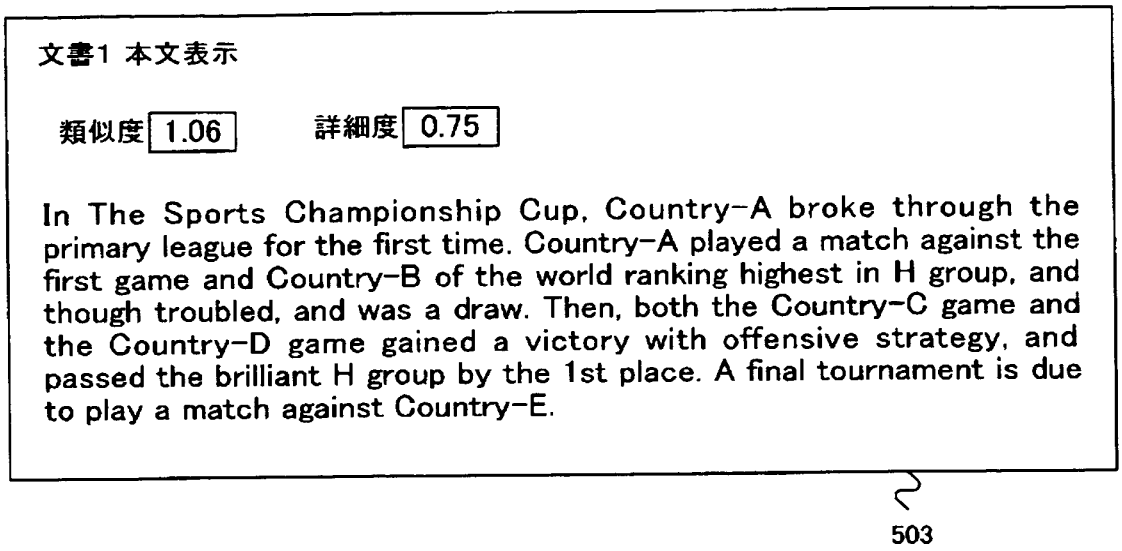
類似度 [0] 以上　詳細度 [0.5] 以上

502

文書ID	類似度	詳細度	見出し
1	1.06	0.75	In The Sports Championship Cup, Country-A …

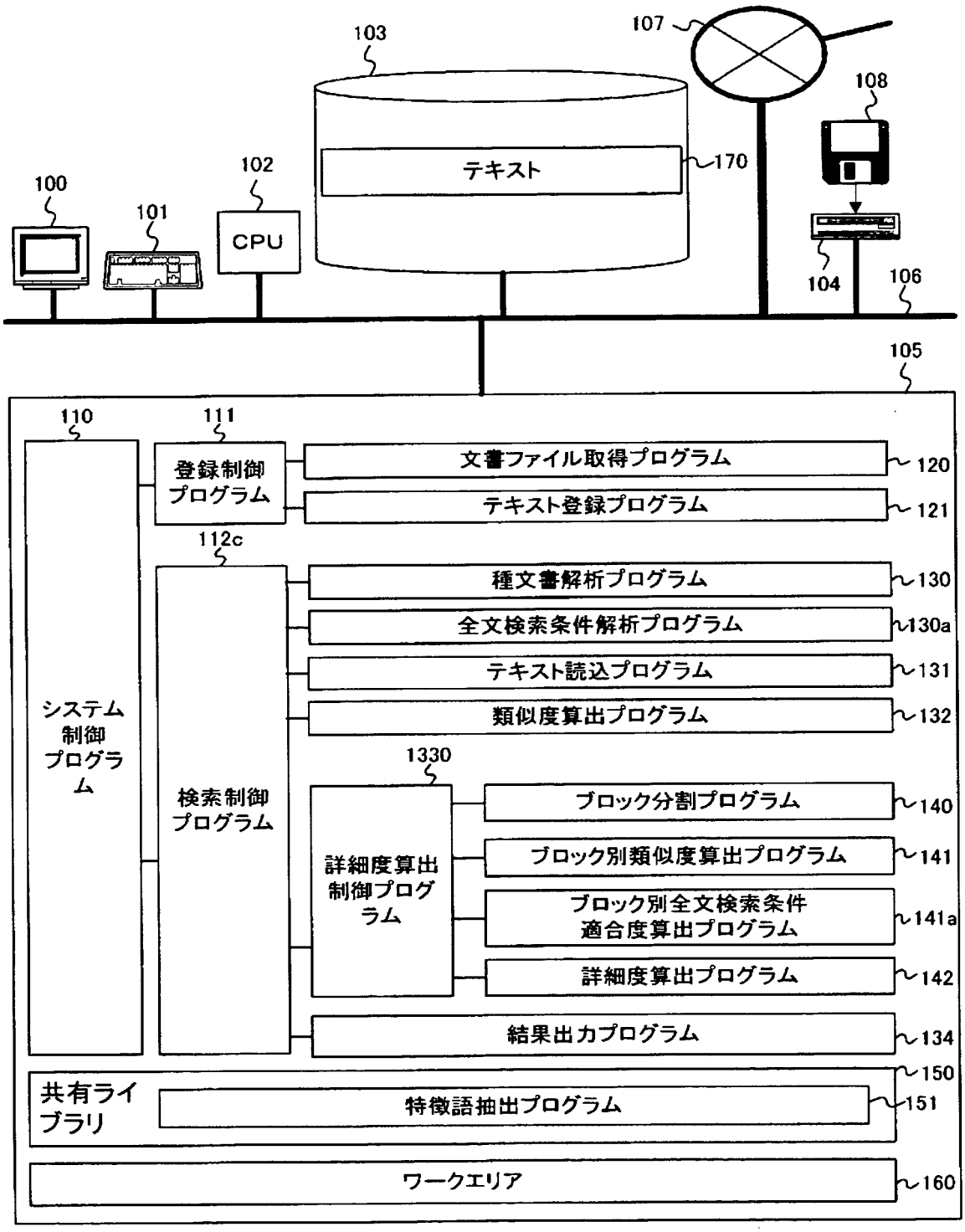
【図８】

図8


文書1 本文表示
類似度 1.06
詳細度 0.75
In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against the first game and Country-B of the world ranking highest in H group, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.
503

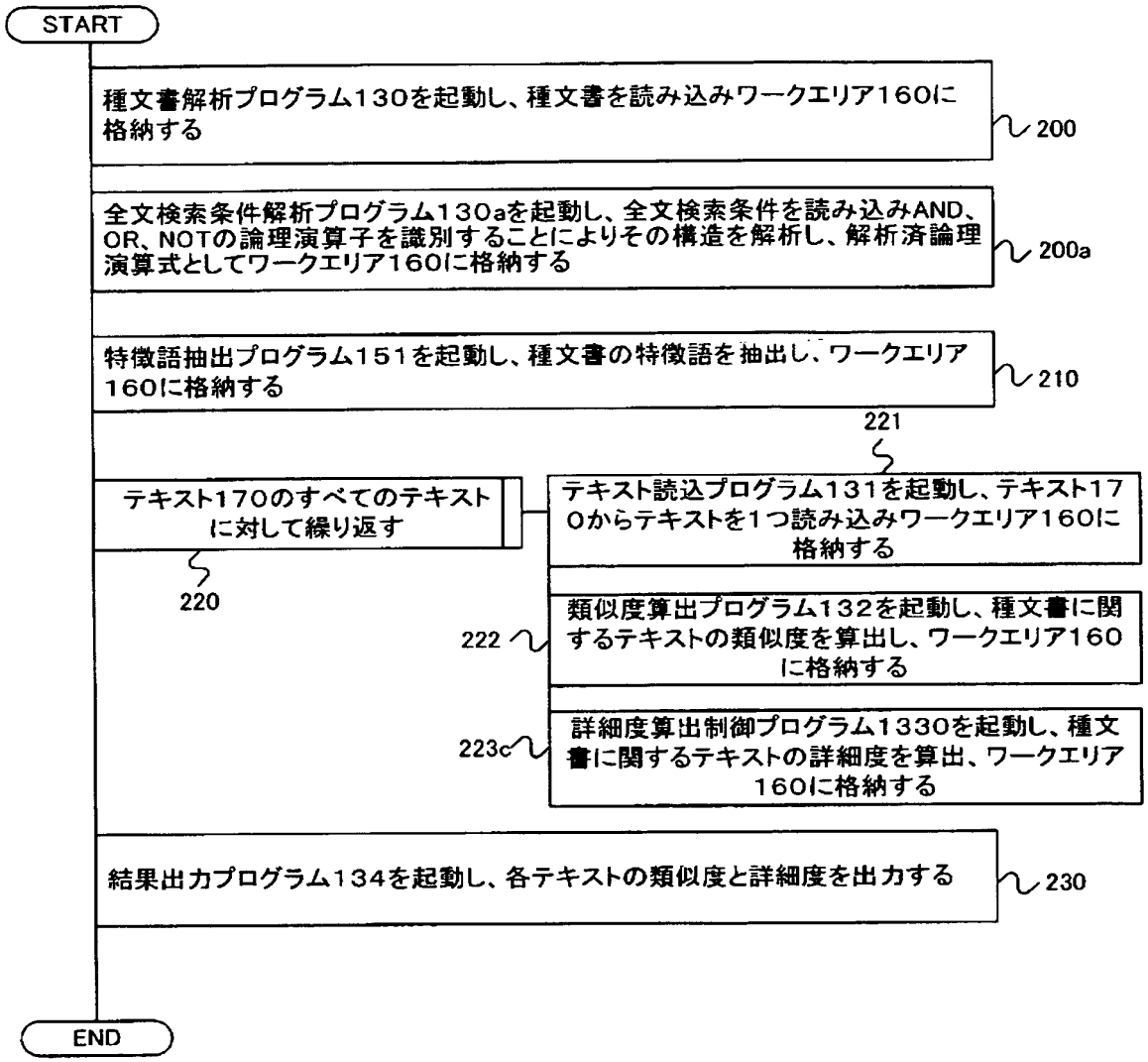
【図９】

図9


103
107
108
テキスト
170
100
101
102
CPU
104
106
105
110
111
登録制御プログラム
文書ファイル取得プログラム
120
テキスト登録プログラム
121
112c
種文書解析プログラム
130
全文検索条件解析プログラム
130a
テキスト読込プログラム
131
類似度算出プログラム
132
システム制御プログラム
検索制御プログラム
1330
詳細度算出制御プログラム
ブロック分割プログラム
140
ブロック別類似度算出プログラム
141
ブロック別全文検索条件適合度算出プログラム
141a
詳細度算出プログラム
142
結果出力プログラム
134
150
共有ライブラリ
特徴語抽出プログラム
151
ワークエリア
160

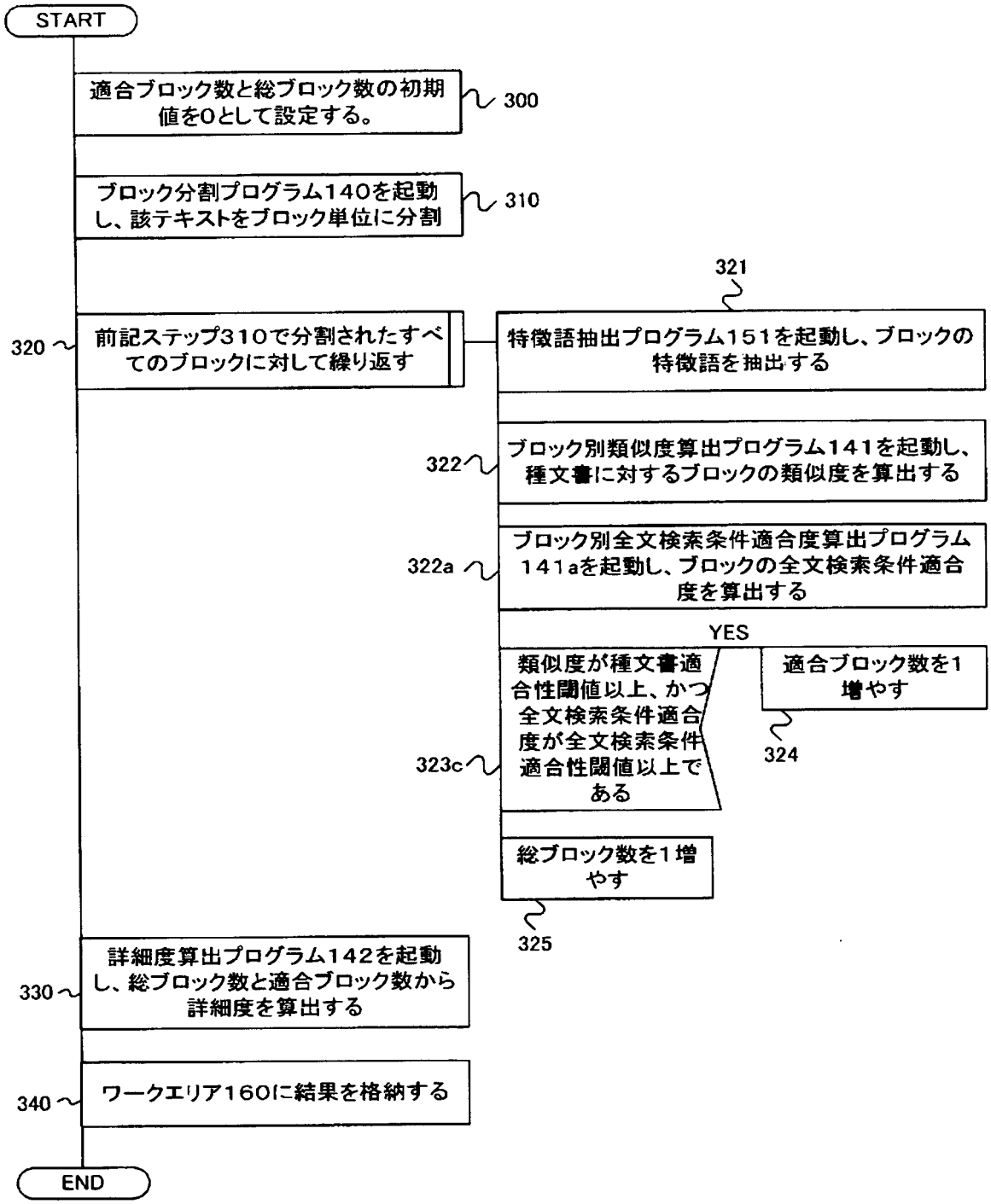
【図１０】

図10


START
種文書解析プログラム130を起動し、種文書を読み込みワークエリア160に格納する
200
全文検索条件解析プログラム130aを起動し、全文検索条件を読み込みAND、OR、NOTの論理演算子を識別することによりその構造を解析し、解析済論理演算式としてワークエリア160に格納する
200a
特徴語抽出プログラム151を起動し、種文書の特徴語を抽出し、ワークエリア160に格納する
210
221
テキスト170のすべてのテキストに対して繰り返す
220
テキスト読込プログラム131を起動し、テキスト170からテキストを1つ読み込みワークエリア160に格納する
222
類似度算出プログラム132を起動し、種文書に関するテキストの類似度を算出し、ワークエリア160に格納する
223c
詳細度算出制御プログラム1330を起動し、種文書に関するテキストの詳細度を算出、ワークエリア160に格納する
結果出力プログラム134を起動し、各テキストの類似度と詳細度を出力する
230
END

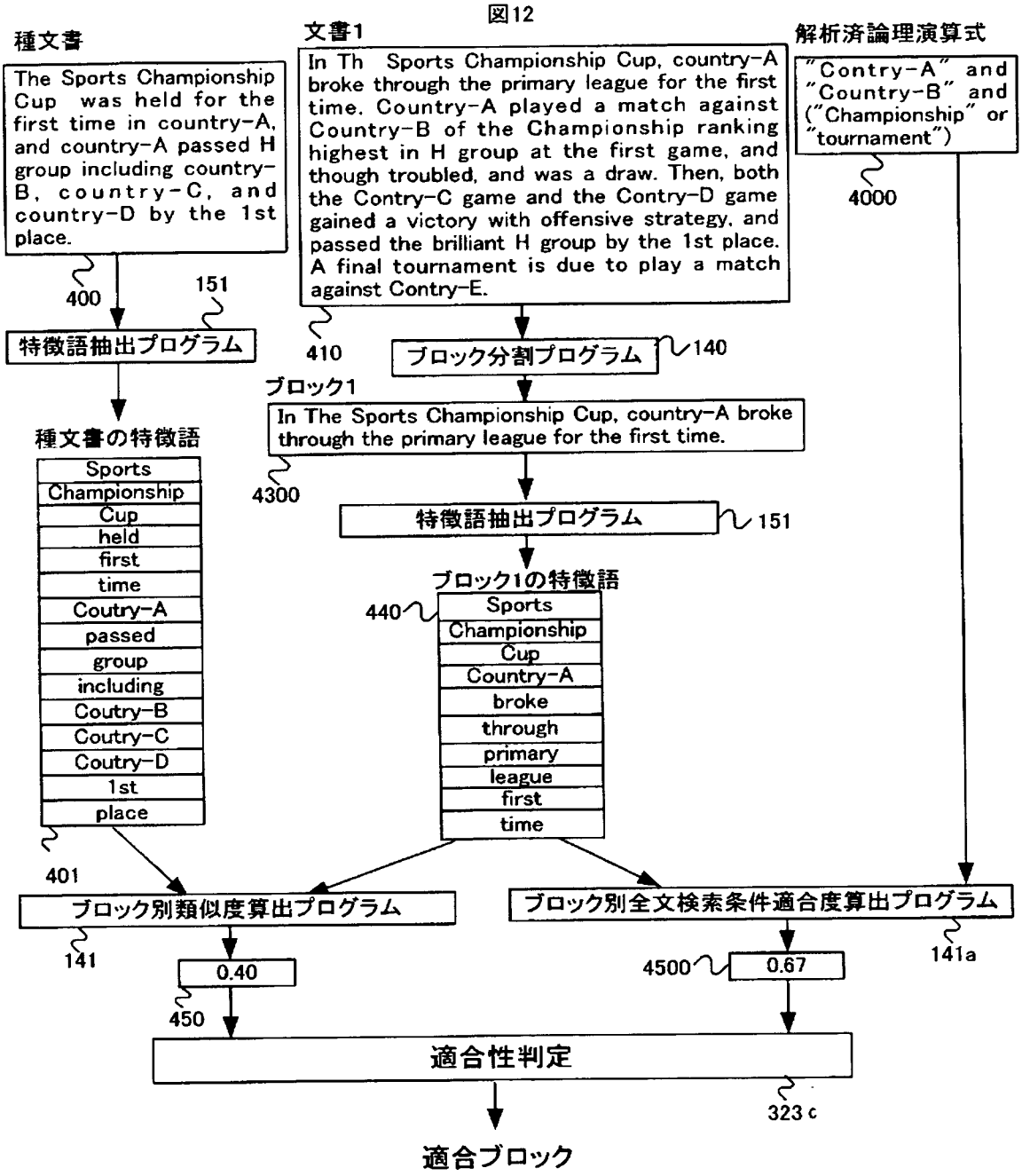
【図１１】

図11


START
適合ブロック数と総ブロック数の初期値を0として設定する。
300
ブロック分割プログラム140を起動し、該テキストをブロック単位に分割
310
321
320
前記ステップ310で分割されたすべてのブロックに対して繰り返す
特徴語抽出プログラム151を起動し、ブロックの特徴語を抽出する
322
ブロック別類似度算出プログラム141を起動し、種文書に対するブロックの類似度を算出する
322a
ブロック別全文検索条件適合度算出プログラム141aを起動し、ブロックの全文検索条件適合度を算出する
YES
類似度が種文書適合性閾値以上、かつ全文検索条件適合度が全文検索条件適合性閾値以上である
323c
適合ブロック数を1増やす
324
総ブロック数を1増やす
325
詳細度算出プログラム142を起動し、総ブロック数と適合ブロック数から詳細度を算出する
330
ワークエリア160に結果を格納する
340
END

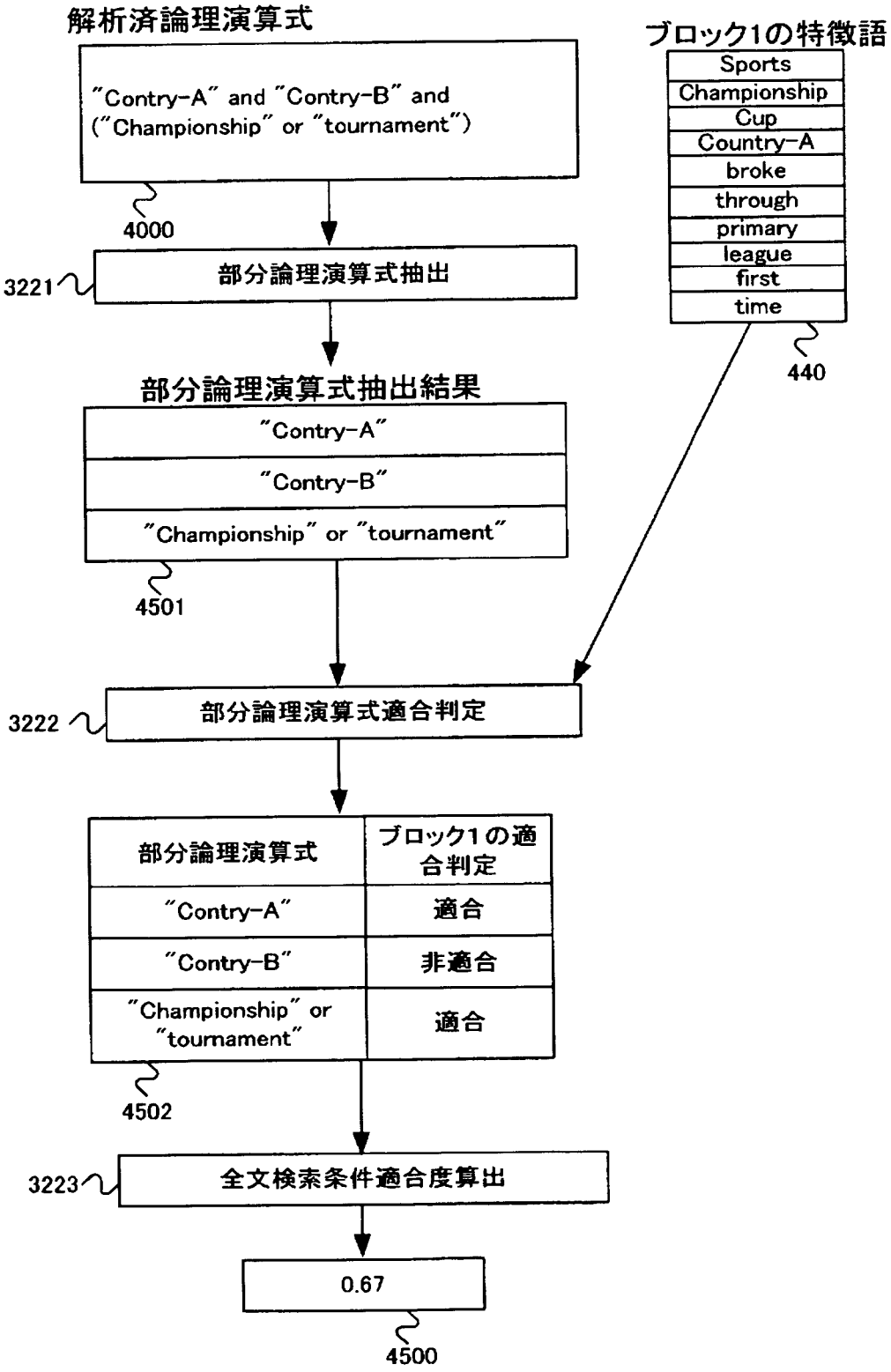
【図１２】

図12


種文書
The Sports Championship Cup was held for the first time in country-A, and country-A passed H group including country-B, country-C, and country-D by the 1st place.
400
151
特徴語抽出プログラム
種文書の特徴語
Sports
Championship
Cup
held
first
time
Coutry-A
passed
group
including
Coutry-B
Coutry-C
Coutry-D
1st
place
401
文書1
In Th Sports Championship Cup, country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Contry-C game and the Contry-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Contry-E.
410
ブロック分割プログラム
140
ブロック1
In The Sports Championship Cup, country-A broke through the primary league for the first time.
4300
特徴語抽出プログラム
151
ブロック1の特徴語
440
Sports
Championship
Cup
Country-A
broke
through
primary
league
first
time
解析済論理演算式
"Contry-A" and "Country-B" and ("Championship" or "tournament")
4000
ブロック別類似度算出プログラム
141
0.40
450
ブロック別全文検索条件適合度算出プログラム
141a
4500
0.67
適合性判定
323 c
適合ブロック

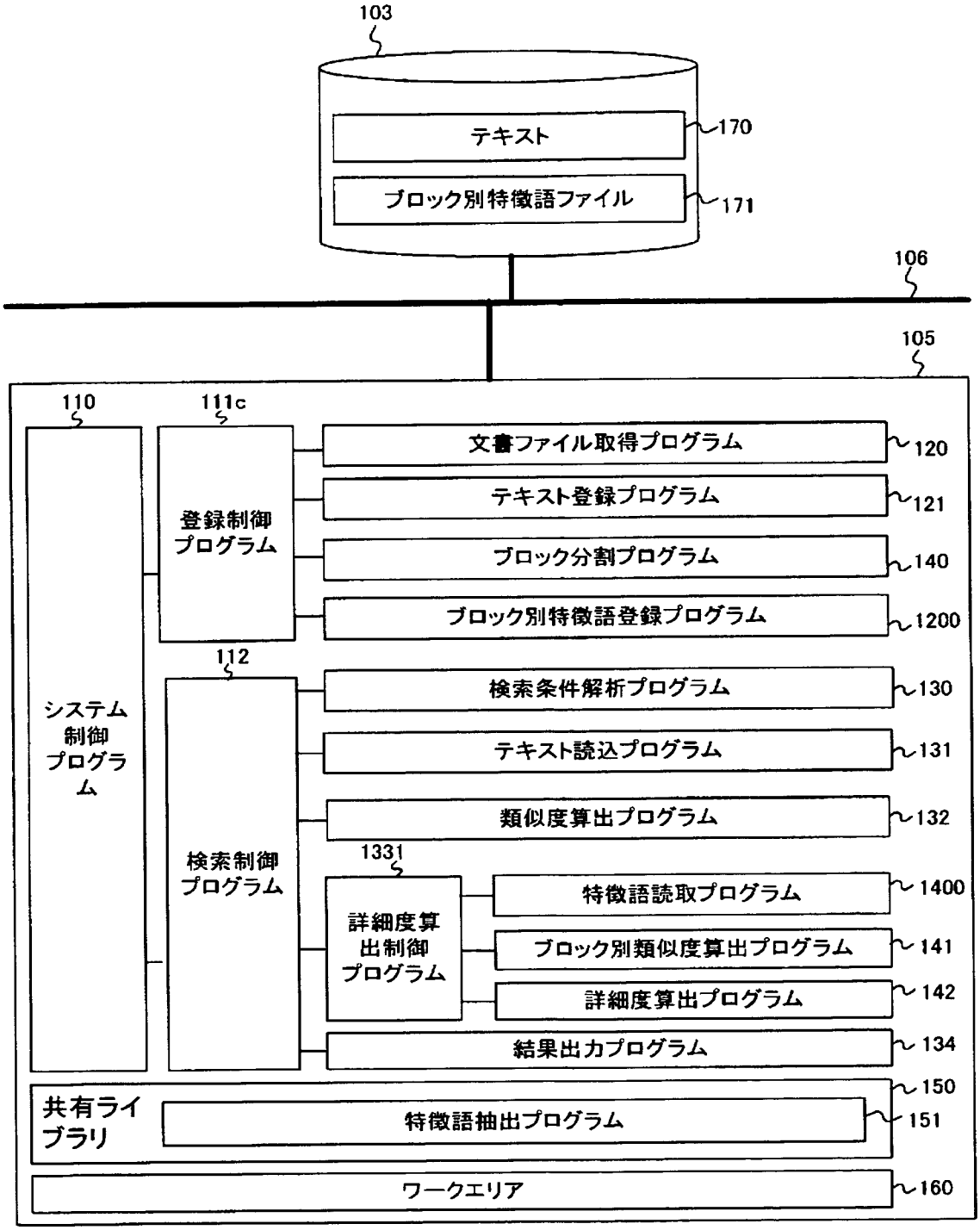
【図１３】

図13


解析済論理演算式
"Contry-A" and "Contry-B" and ("Championship" or "tournament")
4000
3221
部分論理演算式抽出
ブロック1の特徴語
Sports
Championship
Cup
Country-A
broke
through
primary
league
first
time
440
部分論理演算式抽出結果
"Contry-A"
"Contry-B"
"Championship" or "tournament"
4501
3222
部分論理演算式適合判定
部分論理演算式
ブロック1の適合判定
"Contry-A"
適合
"Contry-B"
非適合
"Championship" or "tournament"
適合
4502
3223
全文検索条件適合度算出
0.67
4500

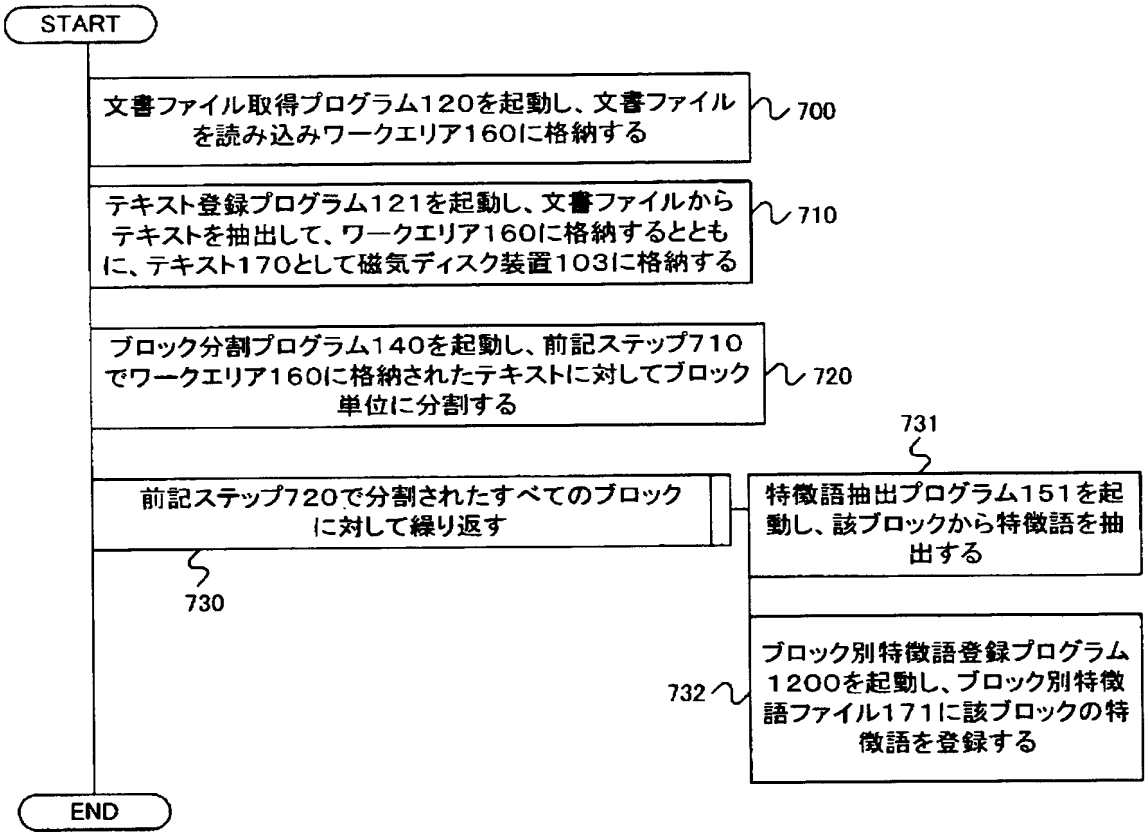
【図１４】

図14


103
テキスト
170
ブロック別特徴語ファイル
171
106
105
110
111c
システム制御プログラム
登録制御プログラム
文書ファイル取得プログラム
120
テキスト登録プログラム
121
ブロック分割プログラム
140
ブロック別特徴語登録プログラム
1200
112
検索制御プログラム
検索条件解析プログラム
130
テキスト読込プログラム
131
類似度算出プログラム
132
1331
詳細度算出制御プログラム
特徴語読取プログラム
1400
ブロック別類似度算出プログラム
141
詳細度算出プログラム
142
結果出力プログラム
134
共有ライブラリ
150
特徴語抽出プログラム
151
ワークエリア
160

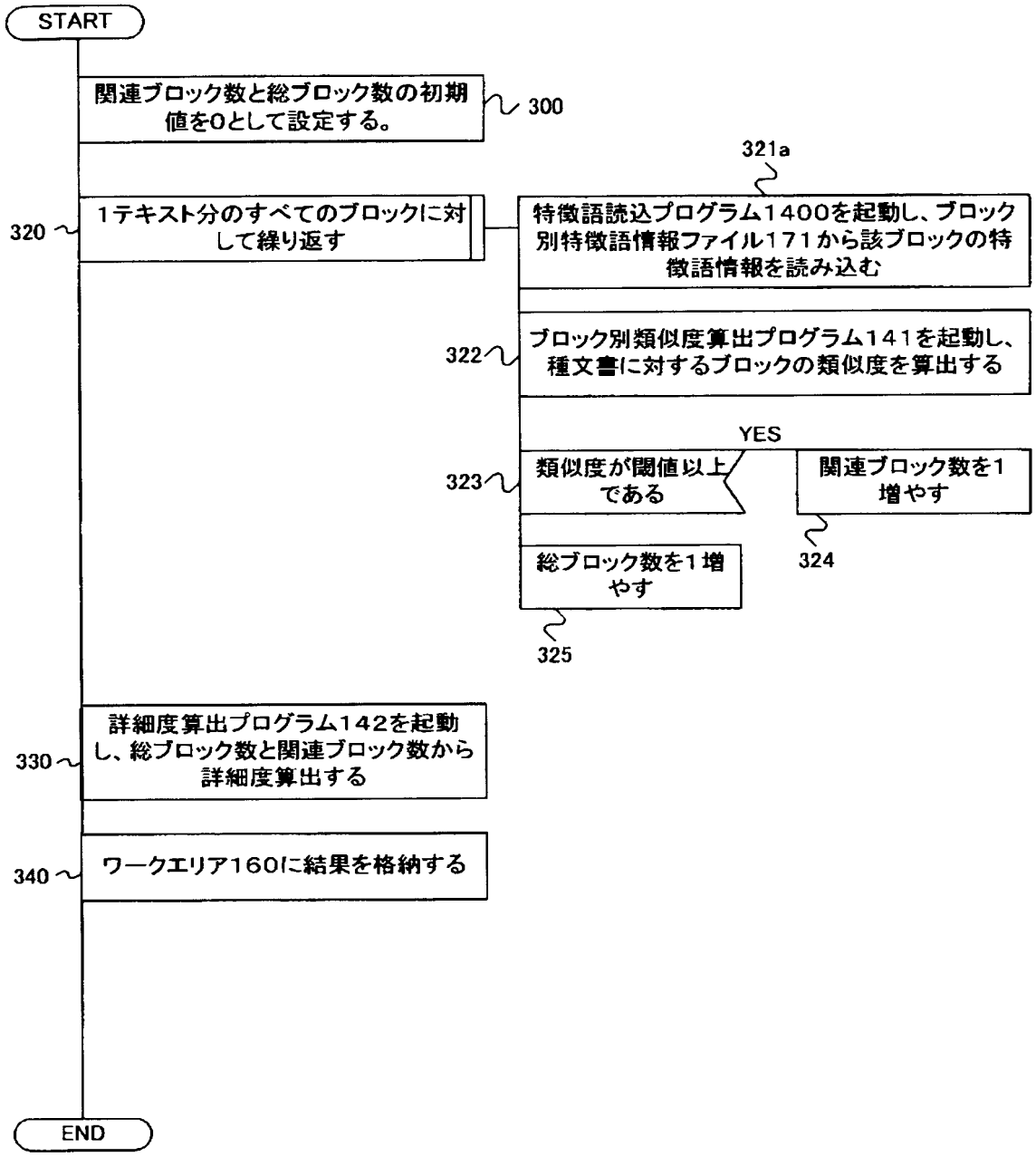
【図１５】

図15


START
文書ファイル取得プログラム120を起動し、文書ファイルを読み込みワークエリア160に格納する
700
テキスト登録プログラム121を起動し、文書ファイルからテキストを抽出して、ワークエリア160に格納するとともに、テキスト170として磁気ディスク装置103に格納する
710
ブロック分割プログラム140を起動し、前記ステップ710でワークエリア160に格納されたテキストに対してブロック単位に分割する
720
731
前記ステップ720で分割されたすべてのブロックに対して繰り返す
730
特徴語抽出プログラム151を起動し、該ブロックから特徴語を抽出する
732
ブロック別特徴語登録プログラム1200を起動し、ブロック別特徴語ファイル171に該ブロックの特徴語を登録する
END

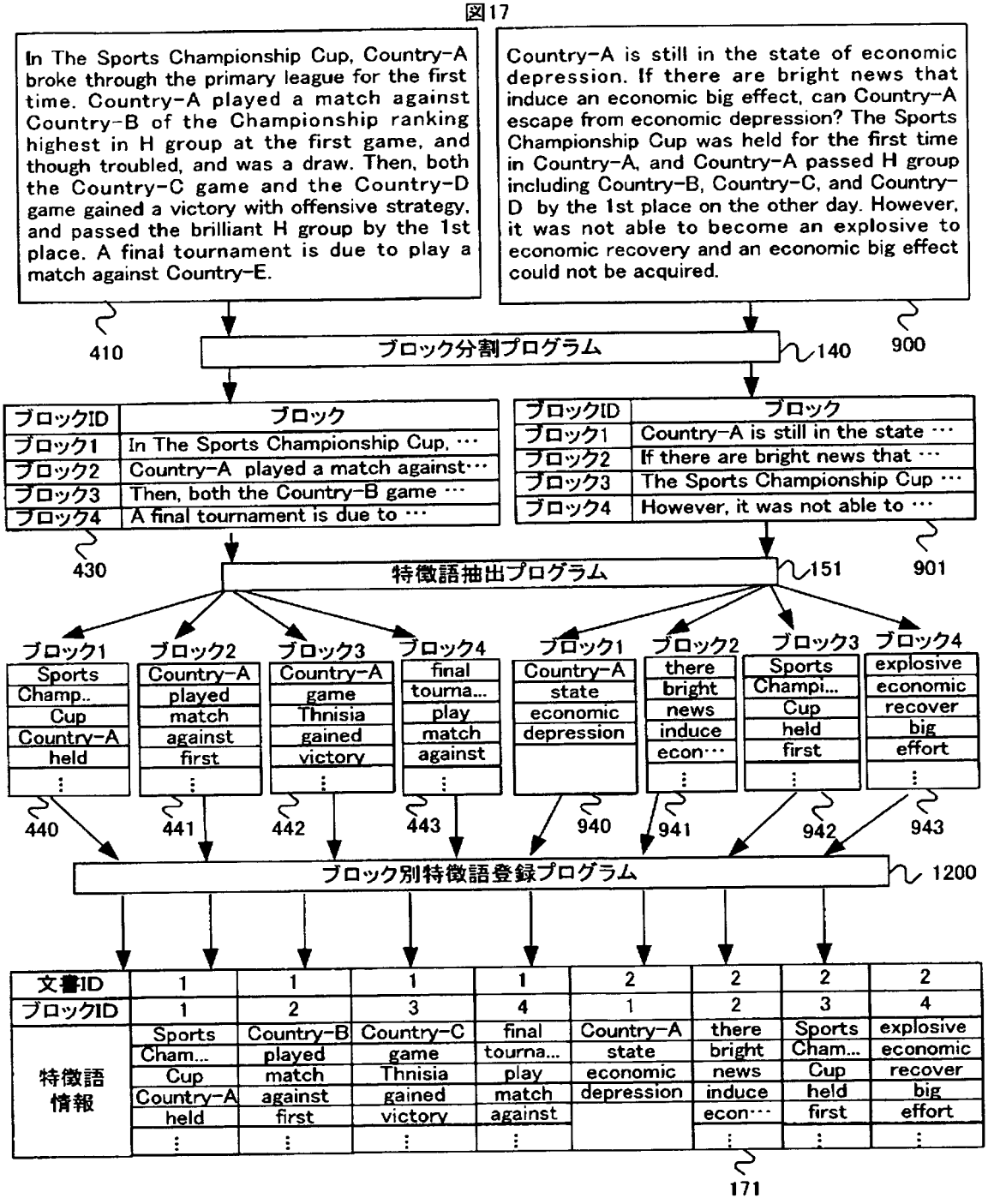
【図１６】

図16


START
関連ブロック数と総ブロック数の初期値を0として設定する。
300
321a
320
1テキスト分のすべてのブロックに対して繰り返す
特徴語読込プログラム1400を起動し、ブロック別特徴語情報ファイル171から該ブロックの特徴語情報を読み込む
322
ブロック別類似度算出プログラム141を起動し、種文書に対するブロックの類似度を算出する
YES
323
類似度が閾値以上である
関連ブロック数を1増やす
324
総ブロック数を1増やす
325
330
詳細度算出プログラム142を起動し、総ブロック数と関連ブロック数から詳細度算出する
340
ワークエリア160に結果を格納する
END

【図１７】

図17


In The Sports Championship Cup, Country-A broke through the primary league for the first time. Country-A played a match against Country-B of the Championship ranking highest in H group at the first game, and though troubled, and was a draw. Then, both the Country-C game and the Country-D game gained a victory with offensive strategy, and passed the brilliant H group by the 1st place. A final tournament is due to play a match against Country-E.
Country-A is still in the state of economic depression. If there are bright news that induce an economic big effect, can Country-A escape from economic depression? The Sports Championship Cup was held for the first time in Country-A, and Country-A passed H group including Country-B, Country-C, and Country-D by the 1st place on the other day. However, it was not able to become an explosive to economic recovery and an economic big effect could not be acquired.
410
ブロック分割プログラム
140
900
ブロックID | ブロック
ブロック1 | In The Sports Championship Cup, ···
ブロック2 | Country-A played a match against···
ブロック3 | Then, both the Country-B game ···
ブロック4 | A final tournament is due to ···
ブロックID | ブロック
ブロック1 | Country-A is still in the state ···
ブロック2 | If there are bright news that ···
ブロック3 | The Sports Championship Cup ···
ブロック4 | However, it was not able to ···
430
特徴語抽出プログラム
151
901
ブロック1 Sports Champ.. Cup Country-A held
ブロック2 Country-A played match against first
ブロック3 Country-A game Thnisia gained victory
ブロック4 final tourna... play match against
ブロック1 Country-A state economic depression
ブロック2 there bright news induce econ···
ブロック3 Sports Champi... Cup held first
ブロック4 explosive economic recover big effort
440
441
442
443
940
941
942
943
ブロック別特徴語登録プログラム
1200
文書ID | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2
ブロックID | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4
特徴語情報 | Sports | Country-B | Country-C | final | Country-A | there | Sports | explosive
Cham... | played | game | tourna... | state | bright | Cham... | economic
Cup | match | Thnisia | play | economic | news | Cup | recover
Country-A | against | gained | match | depression | induce | held | big
held | first | victory | against | | econ··· | first | effort
171

【書類名】　要約書

【要約】

【課題】

文書間の類似性を判断するための指標を算出する類似文書検索方法を提供することにある。

【解決手段】

予め記憶された検索対象の対象文書の中から文書を検索する検索条件として入力された種文書に含まれる文字列を抽出し、対象文書を複数の部分に分割して、分割した対象文書の各部分に含まれる文字列を抽出し、これら文字列を比較して、前記分割された部分ごとに前記種文書に対する類似度を算出するとともに、その類似度と予め定められた閾値とを比較して、分割された各部分が種文書に適合している部分であるか否かの判定結果をもとに、対象文書の前記種文書に対する詳細度を算出する。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００３－０８９６３３
受付番号	５０３００５１０９２２
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　３月３１日

＜認定情報・付加情報＞

【提出日】　平成15年 3月28日

次頁無

特願２００３－０８９６３３

出　願　人　履　歴　情　報

識別番号　　［０００００５１０８］

１．変更年月日　　１９９０年　８月３１日
［変更理由］　　新規登録
住　所　　東京都千代田区神田駿河台４丁目６番地
氏　名　　株式会社日立製作所